A unifying tutorial on Approximate Message Passing

Oliver Y. Feng*, Ramji Venkataramanan†, Cynthia Rush‡ and Richard J. Samworth*
*Statistical Laboratory, University of Cambridge
†Department of Engineering, University of Cambridge
‡Department of Statistics, Columbia University

March 7, 2022

Abstract

Over the last decade or so, Approximate Message Passing (AMP) algorithms have become extremely popular in various structured high-dimensional statistical problems. Although the origins of these techniques can be traced back to notions of belief propagation in the statistical physics literature, our goals in this work are to present the main ideas of AMP from a statistical perspective and to illustrate the power and flexibility of the AMP framework. Along the way, we strengthen and unify many of the results in the existing literature.

1 Introduction

Approximate Message Passing (AMP) refers to a class of iterative algorithms that have been successfully applied to a number of statistical estimation tasks such as linear regression (Donoho et al., 2009; Bayati and Montanari, 2011; Krzakala et al., 2012), generalised linear models (Rangan, 2011; Schniter and Rangan, 2014; Mondelli and Venkataramanan, 2020) and low-rank matrix estimation (Matsushita and Tanaka, 2013; Deshpande and Montanari, 2014; Deshpande et al., 2016; Montanari and Richard, 2016; Kabashima et al., 2016; Lesieur et al., 2017; Rangan and Fletcher, 2018; Montanari and Venkataramanan, 2021). Moreover, these techniques are also popular and practical in a variety of engineering and computer science applications such as imaging (Fletcher and Rangan, 2014; Vila et al., 2015; Metzler et al., 2017), communications (Schniter, 2011; Jeon et al., 2015; Barbier and Krzakala, 2017; Rush et al., 2017) and machine learning (Manoel et al., 2017; El Alaoui et al., 2018; Yang, 2019; Emami et al., 2020; Pandit et al., 2020). AMP algorithms have two features that make them particularly attractive. First, they can easily be tailored to take advantage of prior information on the structure of the signal, such as sparsity or other constraints. Second, under suitable assumptions on a design or data matrix, AMP theory provides precise asymptotic guarantees for statistical procedures in the high-dimensional regime where the ratio of the number of observations n to dimensions p converges to a constant (Bayati and Montanari, 2012; Donoho et al., 2013; Donoho and Montanari, 2016; Sur et al., 2017). More generally, AMP has been used to obtain lower bounds on the estimation error of first-order methods (Celentano et al., 2020). In generalised linear models, low-rank matrix estimation and neural network models, it also plays a fundamental role in understanding the performance gap between information-theoretically optimal and computationally feasible estimators (Aubin et al., 2019, 2020; Barbier et al., 2019; Lelarge and Miolane, 2019; Reeves and Pfister, 2019). In these settings, it is conjectured that AMP achieves the optimal asymptotic estimation error among all polynomial-time algorithms (cf. Celentano and Montanari, 2022).

The purpose of this article is to give a comprehensive and rigorous introduction to what AMP can offer, as well as to unify and formalise the core concepts within the large body of recent work in the area. We mention here that many of the original ideas of AMP were developed in the physics and engineering literature, and involved notions such as ‘loopy belief propagation’ (e.g. Koller and Friedman, 2009, Section 11.3) and the ‘replica method’ (e.g. Tanaka, 2002; Guo and Verdú, 2005; Mézard and Montanari, 2009; Rangan et al., 2009; Krzakala et al., 2012). Our starting point, however, will be an abstract AMP recursion, whose form depends on whether or not the data matrix is symmetric; we will study the symmetric case in detail, and then present the asymmetric version, which can be handled via a reduction argument. The striking and crucial feature of this recursion is that when the dimension is large, the empirical distribution of the coordinates of each iterate is approximately Gaussian, with limiting variance given by a scalar iteration called ‘state evolution’.

Rigorous formulations of the key AMP property are given in Theorems 2.1 and 2.3 (for the symmetric case) and Theorem 2.5 (for the asymmetric case), which can be found in Sections 2.1 and 2.2 respectively. Here, we both strengthen earlier related results, and seek to make the underlying arguments more transparent. These ‘master theorems’, which can be viewed as asymptotic results on Gaussian random matrices, can be adapted to analyse variants of the original AMP recursion that are geared towards more statistical problems. In this aspect, we focus on two canonical statistical settings, namely estimation of low-rank matrices in Section 3, and estimation in generalised linear models (GLMs) in Section 4. The former encompasses Sparse Principal Component Analysis (Jolliffe et al., 2003; Zou et al., 2006; Deshpande and Montanari, 2014; Wang et al., 2016; Gataric et al., 2020), submatrix detection (Ma and Wu, 2015), hidden clique detection (Alon et al., 1998; Deshpande and Montanari, 2015), spectral clustering (von Luxburg, 2007), matrix completion (Candès and Recht, 2009; Zhu et al., 2019), topic modelling (Blei et al., 2003) and collaborative filtering (Su and Khoshgoftaar, 2009). The latter provides a holistic approach to studying a suite of popular modern statistical methods, including penalised M-estimators such as the Lasso (Tibshirani, 1996) and SLOPE (Bogdan et al., 2015), as well as more traditional techniques such as logistic regression. A novel aspect of our presentation in Section 4 is that we formalise the connection between AMP and a broad class of convex optimisation problems, and then show how to systematically derive exact expressions for the asymptotic risk of estimators in GLMs. We expect that our general recipe can be applied to a wider class of GLMs than have been studied in the AMP literature to date.

To preview the statistical content in this tutorial and highlight some recurring themes, we now discuss two prototypical applications of AMP that form the basis of Sections 3 and 4 respectively. First, suppose that we wish to estimate an unknown signal $v \in \mathbb{R}^n$ based on an observation

$$A = \frac{\lambda}{n} v v^\top + W,$$

where $\lambda > 0$ is fixed and $W \in \mathbb{R}^{n \times n}$ is a symmetric Gaussian noise matrix. In this so-called spiked Wigner model (see Section 3.1 and the references therein), a popular and well-studied estimator of v is the leading eigenvector $\hat{\varphi}$ of A, which can be approximated via the power method, with iterates

$$v^{k+1} = \frac{Av^k}{\|Av^k\|}.$$

An AMP algorithm in this context can be interpreted as a generalised power method that produces a sequence of estimates $\hat{v}^k$ of v via iterative updates of the form

$$\hat{v}^k = g_k(v^k), \qquad v^{k+1} = A\hat{v}^k - b_k \hat{v}^{k-1}$$

for $k \in \mathbb{N}_0$, where we emphasise the following two characteristic features:

(i) Each ‘denoising’ function $g_k \colon \mathbb{R} \to \mathbb{R}$ is applied componentwise to vectors, and can be chosen appropriately to exploit different types of prior information about the structure of v (e.g. to encourage $\hat{v}^k$ to be sparse).

(ii) In the ‘memory’ term $-b_k \hat{v}^{k-1}$, which is called an ‘Onsager’ correction in the AMP literature (e.g. Donoho et al., 2009; Bayati and Montanari, 2011), the scalar b_k is defined as a specific function of v^k to ensure that the iterates v^{k+1} have desirable statistical properties; see (3.3) below.

One way to incorporate additional structural information on v into the spiked model is to assume that its entries are drawn independently from some prior distribution π on $\mathbb{R}$; for example, we can enforce sparsity through priors that place strictly positive mass at 0. Then under appropriate conditions, AMP theory guarantees that, for each k, the components of the estimate $\hat{v}^k$ have approximately the same empirical distribution as those of $g_k(\mu_k v + \sigma_k \xi)$; here, $\xi \sim N_n(0, I_n)$ is a 'noise' vector that is independent of the signal $v \in \mathbb{R}^n$, and the 'signal' and 'noise' parameters $\mu_k \in \mathbb{R}$, $\sigma_k > 0$ are determined by a scalar state evolution recursion that depends on (g_k) and the prior distribution π; see (3.6). This distributional characterisation effectively reduces the analysis of the high-dimensional $\hat{v}^k$ to a much simpler univariate denoising problem, where the aim is to reconstruct $V \sim \pi$ based on a single corrupted observation of the form $\mu_k V + \sigma_k G$ with $G \sim N(0, 1)$ representing independent Gaussian noise. The functions g_k can then be chosen in such a way that the 'effective signal-to-noise ratios' $(\mu_k/\sigma_k)^2$ are large and $g_k(\mu_k V + \sigma_k G)$ accurately estimates V. This ensures that the resulting AMP estimates $\hat{v}^k = g_k(v^k)$ have low asymptotic estimation error as $n \to \infty$.

For instance, suppose that the entries of v are drawn uniformly at random from $\{-1, 1\}$. Then provided we initialise the AMP algorithm with $v^0 = \hat{\varphi}$ and $\hat{v}^{-1} = \lambda^{-1}\hat{\varphi}$, where $\|\hat{\varphi}\| = \sqrt{n\lambda^2(\lambda^2 - 1)_+}$, it turns out that the asymptotic mean squared error (MSE) of $\hat{v}^k$ is minimised by choosing g_k to be the function $x \mapsto \tanh(\mu_k x/\sigma_k^2)$; see Section 3.3. Figure 1 illustrates that the limiting MSE of the AMP estimates $\hat{v}^k$ decreases with the iteration number k, and in particular that they improve on the pilot spectral estimator $\hat{v}^{-1}$ (which is agnostic to the structure of v).

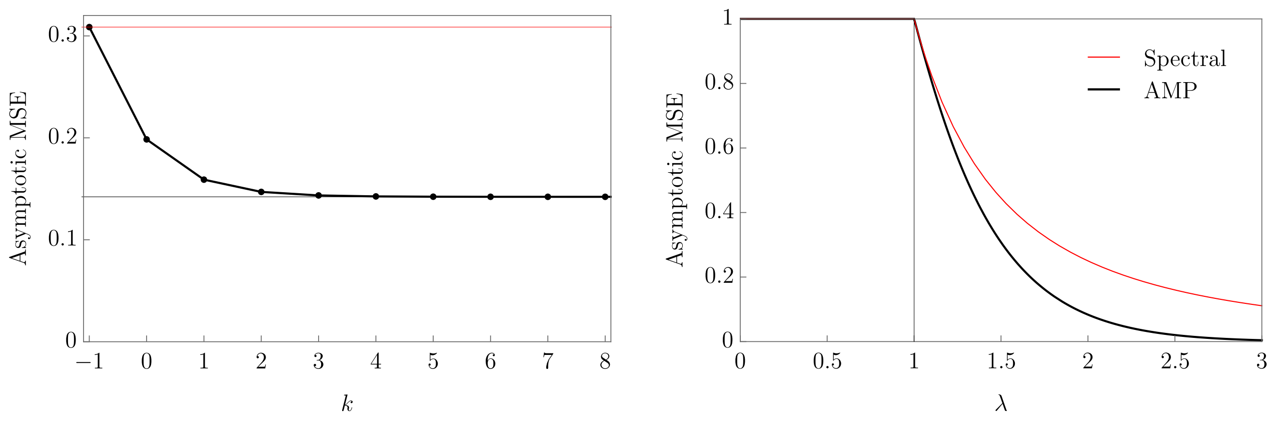

Figure 1: Asymptotic mean-squared error plots for estimation of a signal $v \in \mathbb{R}^n$ with i.i.d. $U\{-1, 1\}$ entries in the rank-one spiked model, based on an AMP algorithm with denoising functions $g_k \colon x \mapsto \tanh(\mu_k x/\sigma_k^2)$ and spectral initialisation ($v^0 = \hat{\varphi}$ and $\hat{v}^{-1} = \lambda^{-1}\hat{\varphi}$ with $\|\hat{\varphi}\| = \sqrt{n\lambda^2(\lambda^2 - 1)_+}$). See Sections 3.2–3.3, where we also discuss how to consistently estimate λ when it is unknown (Remark 3.12).

Left: Plot of $\mathrm{AMSE}_k(\lambda) := \lim_{n\to\infty} \|\hat{v}^k - v\|^2/n$ against the iteration number k for the AMP estimates $\hat{v}^k \equiv \hat{v}^k_\lambda(n)$, when $\lambda = 1.7$. $\mathrm{AMSE}_k(\lambda)$ decreases monotonically to some $\mathrm{AMSE}_\infty(\lambda)$ as $k \to \infty$; see Theorem 3.10(c).

Right: Plots of $\mathrm{AMSE}_{-1}(\lambda) = 1 \wedge \lambda^{-2}$ for the pilot spectral estimator $\hat{v}^{-1}$ and $\mathrm{AMSE}_\infty(\lambda)$ for AMP, with $\lambda \in [0, 3]$. The spectral estimator undergoes the so-called BBP phase transition at $\lambda = 1$; see Section 3.1.

As a second example, consider the linear model $y = X\beta + \varepsilon$, where $\beta \in \mathbb{R}^p$ is the target of inference, $\varepsilon \in \mathbb{R}^n$ is a noise vector, and $X \in \mathbb{R}^{n\times p}$ is a random design matrix with independent $N(0, 1/n)$ entries. In high-dimensional regimes where p is comparable in magnitude to, or even much larger than n, a popular (sparse) estimator is the Lasso (Tibshirani, 1996), which for $\lambda > 0$ is defined by

$$\hat{\beta}^{\mathrm{L},\lambda} \in \operatorname*{argmin}_{\tilde{\beta}\in\mathbb{R}^p} \biggl\{\frac{1}{2}\|y - X\tilde{\beta}\|^2 + \lambda\|\tilde{\beta}\|_1\biggr\}.$$

In the literature on high-dimensional estimation, upper bounds on the prediction and estimation error of the Lasso have been obtained under suitable conditions on the design matrix X, such as the restricted isometry property or compatibility conditions (e.g. Bühlmann and van de Geer, 2011). AMP offers complementary guarantees by providing exact formulae for the asymptotic risk in the 'large system limit' where $n, p \to \infty$ with $n/p \to \delta \in (0, \infty)$, and with the components of β drawn independently from a prior distribution on $\mathbb{R}$. To motivate the form of the AMP algorithm in this setting, first consider the iterative soft thresholding algorithm (ISTA) for solving the Lasso optimisation problem, whose update steps can be written as

$$\hat{r}^k = y - X\hat{\beta}^k, \qquad \hat{\beta}^{k+1} = \mathrm{ST}_{\lambda\eta_k}\big(\hat{\beta}^k + \eta_k X^\top \hat{r}^k\big) \qquad \text{for } k \in \mathbb{N}_0; \tag{1.1}$$

here, $\hat{r}^k$ is the current residual, $\eta_k > 0$ is a deterministic step size, and for $t > 0$, the soft-thresholding function $\mathrm{ST}_t \colon w \mapsto \mathrm{sgn}(w)(|w| - t)_+$ is applied componentwise to vectors. This is an instance of the general-purpose proximal gradient method (Parikh and Boyd, 2013, Sections 4.2 and 4.3). An 'accelerated' version of (1.1) called FISTA (Beck and Teboulle, 2009) bears a closer resemblance to an AMP algorithm, where the iterates of the latter are given by

$$\hat{r}^k = y - X\hat{\beta}^k + \frac{\|\hat{\beta}^k\|_0}{n}\,\hat{r}^{k-1}, \qquad \hat{\beta}^{k+1} = \mathrm{ST}_{t_{k+1}}\big(\hat{\beta}^k + X^\top \hat{r}^k\big) \qquad \text{for } k \in \mathbb{N}_0. \tag{1.2}$$

Here, each $t_k > 0$ is a deterministic threshold and $\|\hat{\beta}^k\|_0$ denotes the number of non-zero entries of $\hat{\beta}^k \in \mathbb{R}^p$. By comparison with (1.1), we observe that $\hat{r}^k$ in (1.2) is a corrected residual, whose definition includes an additional memory term that is crucial for ensuring that the empirical distribution of the iterates can be characterised exactly. Indeed, for each fixed $k \in \mathbb{N}$, the entries of the AMP estimate $\hat{\beta}^k$ of β have approximately the same empirical distribution as those of $\mathrm{ST}_{t_k}(\beta + \sigma_k \xi)$ when p is large; here $\xi \sim N_p(0, I_p)$ is a noise vector that is independent of β, the noise level $\sigma_k > 0$ is determined by the state evolution recursion defined in (4.41) below, and the scalar denoising function ST_{t_k} induces sparsity.

Bayati and Montanari (2012) proved that in the asymptotic regime above, the AMP iterates $(\hat{r}^k, \hat{\beta}^k)$ converge in a suitable sense to a fixed point $(\hat{r}^*, \hat{\beta}^*)$, and a key property of (1.2) is that for any such fixed point, $\hat{\beta}^*$ is a Lasso solution; see (4.42) below. It follows that the performance of the Lasso is precisely characterised by a fixed point of the state evolution recursion (4.41); see Theorem 4.5. Since the above properties are proved under a Gaussian design, the main utility of AMP in this setting is not so much as an efficient Lasso computational algorithm, but rather as a device for gaining insight into the statistical properties of the estimator. In Section 4, the above theory is developed as part of an overarching AMP framework for linear models and generalised linear models (GLMs).

Note that in both of the examples above, the limiting empirical distributions of the entries of the AMP iterates can be decomposed into independent 'signal' and 'noise' components, and the effective signal strength and noise level are determined by a state evolution recursion. In Sections 3 and 4, we show how to derive these asymptotic guarantees by applying the master theorems in Section 2 to suitable abstract recursions, which track the evolution of the asymptotically Gaussian 'noise' components of the AMP iterates. We discuss various extensions in Section 5, and provide proofs in the Appendix (Section 6), with supplementary mathematical background deferred to Section 7. As a guide to the reader, we remark that rigorous formulations of the results in this monograph require a number of technical conditions. While we take care to state these precisely, and discuss them at appropriate places, we emphasise that these should generally be regarded as mild. We therefore recommend that the reader initially focuses on the main conclusions of the results.

The statistical roots of AMP lie in compressed sensing (Donoho et al., 2009, 2013). A reader approaching the subject from this perspective can consult Montanari (2012), Tramel et al. (2014) and Schniter (2020) for accessible expositions of the motivating ideas and the connections with message passing algorithms on dense graphs. Alternatively, for comprehensive reviews of AMP from a statistical physics perspective, see Zdeborová and Krzakala (2016), Krzakala et al. (2012) and Lesieur et al. (2017).

In spin glass theory, an AMP algorithm was proposed as an iterative scheme for solving the Thouless–Anderson–Palmer (TAP) equations corresponding to a Sherrington–Kirkpatrick model with specific parameters (Mézard et al., 1987; Mézard and Montanari, 2009; Talagrand, 2011; Bolthausen, 2014).

The estimation problem here is equivalent to one of reconstructing a symmetric rank-one matrix in a Gaussian spiked model. Bolthausen (2014) proved a rigorous state evolution result for AMP in this specific setting, by introducing a conditioning argument that became an essential ingredient in subsequent analyses of AMP (Bayati and Montanari, 2011; Javanmard and Montanari, 2013; Berthier et al., 2020; Fan, 2022). See Section 6.2 for a detailed discussion of this proof technique.

In this article, we restrict our focus to AMP recursions in which the random matrices are Gaussian. However, as we discuss in Section 5, several recent works have extended AMP and its state evolution recursion to more general non-Gaussian settings. For matrices with independent sub-Gaussian entries, results on the 'universality' of AMP were first established by Bayati et al. (2015) and later in greater generality by Chen and Lam (2021). In addition, to accommodate the class of rotationally invariant random matrices, a number of extensions of the original AMP framework have recently been proposed, including Orthogonal AMP (Ma and Ping, 2017; Takeuchi, 2020) and Vector AMP (Schniter et al., 2016; Rangan et al., 2019b), as well as the general iterative schemes of Opper et al. (2016), Çakmak and Opper (2019) and Fan (2022). Some of these are closely related to expectation propagation (Opper and Winther, 2005; Kabashima and Vehkaperä, 2014). In all of the above variants of AMP, the recursion is tailored to the spectrum of the random matrix.

1.1 Notation and preliminaries

Here, we introduce some notation used throughout this tutorial, and present basic properties of Wasserstein distances, pseudo-Lipschitz functions, as well as the complete convergence of random sequences.

General notation: For $n \in \mathbb{N}$, let $e_1, \dots, e_n$ be the standard basis vectors in $\mathbb{R}^n$. For $r \in [1, \infty]$, we write $\|x\|_r$ for the ℓ_r norm of $x \equiv (x_1, \dots, x_n) \in \mathbb{R}^n$, so that $\|x\|_r = (\sum_{i=1}^n |x_i|^r)^{1/r}$ when $r \in [1, \infty)$ and $\|x\|_\infty = \max_{1 \le i \le n} |x_i|$. We also define $\|x\|_{n,r} := n^{-1/r}\|x\|_r = (n^{-1}\sum_{i=1}^n |x_i|^r)^{1/r}$ for $r \in (1, \infty)$. Let $\langle \cdot\,, \cdot \rangle$ and $\|\cdot\| := \|\cdot\|_2$ be the standard Euclidean inner product and norm on $\mathbb{R}^n$ respectively, and define $\langle \cdot\,, \cdot \rangle_n$ to be the scaled Euclidean inner product on $\mathbb{R}^n$ given by $\langle x, y \rangle_n := n^{-1}\langle x, y \rangle$ for $x, y \in \mathbb{R}^n$, which induces the norm $\|\cdot\|_n := \|\cdot\|_{n,2}$. We denote by $\mathbf{1}_n := (1, \dots, 1) \in \mathbb{R}^n$ the all-ones vector and write $\langle x \rangle_n := \langle x, \mathbf{1}_n \rangle_n = n^{-1}\sum_{i=1}^n x_i$ for each $x \in \mathbb{R}^n$.

For $D \in \mathbb{N}$ and $x^1, \dots, x^D \in \mathbb{R}^n$, we denote by $\nu_n(x^1, \dots, x^D) := n^{-1}\sum_{i=1}^n \delta_{(x_i^1, \dots, x_i^D)}$ the joint empirical distribution of their components, and for a function $f\colon \mathbb{R}^D \to \mathbb{R}$, write $f(x^1, \dots, x^D) := \bigl(f(x_i^1, \dots, x_i^D) : 1 \le i \le n\bigr) \in \mathbb{R}^n$ for the row-wise application of f to $(x^1 \;\cdots\; x^D)$.

By a *Euclidean space* $(E, \|\cdot\|_E)$ we mean a finite-dimensional inner product space over $\mathbb{R}$, equipped with the norm induced by its inner product; examples include $(\mathbb{R}^n, \|\cdot\|)$ for $n \in \mathbb{N}$ and $(\mathbb{R}^{k \times \ell}, \|\cdot\|_{\mathrm{F}})$ for $k, \ell \in \mathbb{N}$, where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm induced by the trace inner product $(A, B) \mapsto \mathrm{tr}(A^\top B)$.

Gaussian orthogonal ensemble: We write $W \sim \mathrm{GOE}(n)$ if $W = (W_{ij})_{1 \le i,j \le n}$ takes values in the space of all symmetric $n \times n$ matrices, and has the property that $(W_{ij})_{1 \le i \le j \le n}$ are independent, with $W_{ij} \sim N(0, 1/n)$ for $1 \le i < j \le n$ and $W_{ii} \sim N(0, 2/n)$ for $i = 1, \dots, n$. Writing $\mathbb{O}_n$ for the set of all $n \times n$ orthogonal matrices, we note the orthogonal invariance property of the $\mathrm{GOE}(n)$ distribution: if $Q \in \mathbb{O}_n$ and $W \sim \mathrm{GOE}(n)$, then $Q^\top W Q \sim \mathrm{GOE}(n)$.

Complete convergence of random sequences: The asymptotic results below are formulated in terms of the notion of *complete convergence* (e.g. Hsu and Robbins, 1947; Serfling, 1980, Chapter 1.3). This is a stronger mode of stochastic convergence than almost sure convergence, and is denoted throughout using the symbol $\xrightarrow{c}$. In Definition 1.1 and Proposition 1.2 below, we give two equivalent characterisations of complete convergence and introduce some associated stochastic O symbols.

Definition 1.1. *Let (X_n) be a sequence of random elements taking values in a Euclidean space $(E, \|\cdot\|_E)$. We say that X_n* converges completely *to a deterministic limit $x \in E$, and write $X_n \xrightarrow{c} x$ or* $\text{c-lim}_{n\to\infty} X_n = x$, *if $Y_n \to x$ almost surely for any sequence of E-valued random elements (Y_n) with $Y_n \overset{d}{=} X_n$ for all n.*

We write $X_n = o_c(1)$ if $X_n \xrightarrow{c} 0$, and write $X_n = O_c(1)$ if $Y_n = O_{a.s.}(1)$ (i.e. $\limsup_{n\to\infty} \|Y_n\|_E < \infty$ almost surely) for any sequence of E-valued random elements (Y_n) with $Y_n \overset{d}{=} X_n$ for all n.

Proposition 1.2. *For a sequence (X_n) of random elements taking values in a Euclidean space $(E, \|\cdot\|_E)$, we have*

(a) $X_n = o_c(1)$ if and only if $\sum_n \mathbb{P}(\|X_n\|_E > \varepsilon) < \infty$ for all $\varepsilon > 0$;

(b) $X_n = O_c(1)$ if and only if there exists $C > 0$ such that $\sum_n \mathbb{P}(\|X_n\|_E > C) < \infty$.

For a deterministic $x \in E$, we see that $X_n \xrightarrow{c} x$ if and only if $\sum_n \mathbb{P}(\|X_n - x\|_E > \varepsilon) < \infty$ for all $\varepsilon > 0$. Moreover, if $X_n \xrightarrow{c} x$, then $X_n = O_c(1)$. The proof of Proposition 1.2, along with various other properties of complete convergence and a calculus for $o_c(1)$ and $O_c(1)$ notation, is given in Section 7.1; see also Remark 6.1.

Wasserstein distances and pseudo-Lipschitz functions: For $D \in \mathbb{N}$ and $r \in [1, \infty)$, we write $\mathcal{P}(r) \equiv \mathcal{P}_D(r)$ for the set of all Borel probability measures P on $\mathbb{R}^D$ with $\int_{\mathbb{R}^D} \|x\|^r \, dP(x) < \infty$. For $P, Q \in \mathcal{P}_D(r)$, the r-*Wasserstein distance* between P and Q is defined by

$$d_r(P, Q) := \inf_{(X,Y)} \mathbb{E}(\|X - Y\|^r)^{1/r},$$

where the infimum is taken over all pairs of random vectors (X, Y) defined on a common probability space with $X \sim P$ and $Y \sim Q$. For $P, P_1, P_2, \ldots \in \mathcal{P}_D(r)$, we have $d_r(P_n, P) \to 0$ if and only if both $\int_{\mathbb{R}^D} \|x\|^r \, dP_n(x) \to \int_{\mathbb{R}^D} \|x\|^r \, dP(x)$ and $P_n \to P$ weakly (e.g. Villani, 2003, Theorem 7.12). Furthermore, for $L > 0$, we write $\mathrm{PL}_D(r, L)$ for the set of functions $\psi \colon \mathbb{R}^D \to \mathbb{R}$ such that

$$|\psi(x) - \psi(y)| \le L\|x - y\| \left(1 + \|x\|^{r-1} + \|y\|^{r-1}\right) \tag{1.3}$$

for all $x, y \in \mathbb{R}^D$, and denote by $\mathrm{PL}_D(r) := \bigcup_{L>0} \mathrm{PL}_D(r, L)$ the class of *pseudo-Lipschitz functions* $f \colon \mathbb{R}^D \to \mathbb{R}$ *of order* r. Note that $\mathrm{PL}_D(1, L)$ is precisely the class of all $(3L)$-Lipschitz functions on $\mathbb{R}^D$, and that $\mathrm{PL}_D(s) \subseteq \mathrm{PL}_D(r)$ for any $1 \le s \le r$. Moreover, for any probability measure $P \in \mathcal{P}_D(r)$, we have $\left|\int_{\mathbb{R}^d} \psi \, dP\right| \le L \int_{\mathbb{R}^D} (\|x\| + \|x\|^r) \, dP(x) + |\psi(0)| < \infty$ for all $\psi \in \mathrm{PL}_D(r, L)$. Now for $P, Q \in \mathcal{P}_D(r)$, we define

$$\widetilde{d}_r(P, Q) := \sup_{\psi \in \mathrm{PL}_D(r,1)} \left| \int_{\mathbb{R}^D} \psi \, dP - \int_{\mathbb{R}^D} \psi \, dQ \right|. \tag{1.4}$$

In Section 7.4, we show (among other things) that $\widetilde{d}_r, d_r$ are metrics on $\mathcal{P}_D(r)$ that induce the same topology (Remark 7.18).

2 Master theorems for abstract AMP recursions

2.1 Symmetric AMP

In this subsection, we present an abstract AMP recursion that was first studied by Bolthausen (2014) in a special case[*], and subsequently by Bayati and Montanari (2011, Section 4) and Javanmard and Montanari (2013) in greater generality. Let $(f_k)_{k=0}^\infty$ be a sequence of Lipschitz functions $f_k \colon \mathbb{R}^2 \to \mathbb{R}$, and for $n \in \mathbb{N}$, let $W \equiv W(n) \in \mathbb{R}^{n \times n}$ be a symmetric matrix and $\gamma \equiv \gamma(n) \in \mathbb{R}^n$ be a vector of auxiliary information. Given $m^{-1} \equiv m^{-1}(n) := 0 \in \mathbb{R}^n$ and an initialiser $h^0 \equiv h^0(n) \in \mathbb{R}^n$, recursively define $m^k \equiv m^k(n) \in \mathbb{R}^n$, $b_k \equiv b_k(n) \in \mathbb{R}$ and $h^{k+1} \equiv h^{k+1}(n) \in \mathbb{R}^n$ by

$$m^k := f_k(h^k, \gamma), \qquad b_k := \langle f_k'(h^k, \gamma) \rangle_n = \frac{1}{n} \sum_{i=1}^n f_k'(h_i^k, \gamma_i), \qquad h^{k+1} := Wm^k - b_k m^{k-1} \tag{2.1}$$

[*] In a 2009 workshop, Bolthausen presented his analysis of AMP for the TAP equations, which inspired the work of Bayati and Montanari (2011); see Section 3 of the latter.

for $k \in \mathbb{N}_0$. Here, $f_k' \colon \mathbb{R}^2 \to \mathbb{R}$ is a bounded, Borel measurable function that agrees with the partial derivative of f_k with respect to its first argument, wherever the latter is defined. Note that for each $y \in \mathbb{R}$, the Lipschitz function $x \mapsto f_k(x, y)$ is differentiable Lebesgue almost everywhere (e.g. Federer, 1996, Theorem 3.1.6) with weak derivative $x \mapsto f_k'(x, y)$.

In its generic form, (2.1) is not intended for use as an algorithm to solve any particular estimation problem, but for the following reasons, it underpins the statistical framework for AMP:

(i) *State evolution characterisation of limiting Gaussian distributions*: In an asymptotic regime where conditions (A0)–(A5) below are satisfied (in particular where (A0) requires W to be Gaussian), the key mathematical property of (2.1) is given by (2.3) below: for fixed $k \in \mathbb{N}$, the empirical distributions of the components of $h^k \equiv h^k(n)$ converge completely in Wasserstein distance to a Gaussian limit $N(0, \tau_k^2)$ as $n \to \infty$. The variances τ_k^2 are determined by the state evolution recursion (2.2) below, which depends on the choice of Lipschitz functions $(f_k : k \in \mathbb{N}_0)$. As we will discuss later in this subsection, the so-called Onsager correction term $-b_k m^{k-1}$ plays a pivotal role in ensuring that the asymptotic distributions are indeed Gaussian.

(ii) *Basis for the construction and analysis of AMP algorithms*: In statistical settings, (2.1) cannot be used as a practical procedure when γ and/or W are unobservable; for example, in Section 3 on low-rank matrix estimation, γ represents the unknown target of inference and W is a noise matrix. Instead, one can replace γ and/or W in (2.1) with observed quantities to design an AMP algorithm that produces a sequence of valid estimates of γ; see (3.3) for instance. To analyse the statistical performance of these AMP estimates, it is helpful to be able to recast the algorithm as an abstract recursion of the form (2.1), because its asymptotic characterisation yields exact expressions for the limiting estimation error in terms of the state evolution parameters. Moreover, through judicious choices of the Lipschitz functions f_k, the AMP estimates can be tailored to different types of prior information about the structure of γ.

(iii) *Precursor to other abstract AMP recursions*: By generalising and transforming (2.1), we can obtain state evolution descriptions of the limiting behaviour in a number of related abstract AMP iterations, including those in which the input matrix need not be symmetric (Section 2.2) and/or the iterates themselves are matrices rather than vectors (Section 6.7). These facilitate the analysis of a wider class of AMP algorithms that are not covered directly by (2.1) alone; see for example Section 4 on GAMP.

We will now formalise point (i) above through Theorem 2.1 below. More generally, in Theorem 2.3, we will establish the Wasserstein limit of the joint empirical distributions of the components of $h^1, \dots, h^k, \gamma \in \mathbb{R}^n$ for each fixed k as $n \to \infty$. In view of (ii) and (iii), we will refer to these results as 'master theorems' for symmetric AMP.

We will consider a probabilistic setup where for each $n \in \mathbb{N}$, we have an AMP recursion (2.1) based on a random triple $(m^0, \gamma, W) \equiv \big(m^0(n), \gamma(n), W(n)\big)$ such that

(A0) $W \equiv W(n) \sim \mathrm{GOE}(n)$ and is independent of $(m^0, \gamma) \equiv \big(m^0(n), \gamma(n)\big)$.

Recalling the concepts and definitions from Section 1.1, we assume that for some $r \in [2, \infty)$ and $\tau_1 \in (0, \infty)$, the inputs to (2.1) also satisfy the following conditions as $n \to \infty$:

(A1) There exists a probability distribution $\pi \in \mathcal{P}_1(r)$ such that the empirical distribution $\nu_n(\gamma)$ of the components of $\gamma \equiv \gamma(n)$ satisfies $d_r\big(\nu_n(\gamma), \pi\big) \xrightarrow{c} 0$.

(A2) $\|m^0\|_n \equiv (n^{-1} \sum_{i=1}^n |m_i^0|^2)^{1/2} \xrightarrow{c} \tau_1$ and $\|m^0\|_{n,r} \equiv (n^{-1} \sum_{i=1}^n |m_i^0|^r)^{1/r} = O_c(1)$.

(A3) There exists a Lipschitz $F_0 \colon \mathbb{R} \to \mathbb{R}$ such that taking $\bar{\gamma} \sim \pi$, we have $\mathbb{E}\big(F_0(\bar{\gamma})^2\big) \leq \tau_1^2$ and $\langle m^0, \phi(\gamma) \rangle_n = n^{-1} \sum_{i=1}^n f_0(h_i^0, \gamma_i)\, \phi(\gamma_i) \xrightarrow{c} \mathbb{E}\big(F_0(\bar{\gamma})\phi(\bar{\gamma})\big)$ for all Lipschitz $\phi \colon \mathbb{R} \to \mathbb{R}$.

(A1) holds if for each n, the entries of $\gamma \equiv \gamma(n)$ are drawn independently from a distribution π on $\mathbb{R}$ with a finite r^{th} moment. In general, π can be thought of as a 'limiting prior distribution' in statistical applications. (A2) includes a boundedness assumption on the empirical r^{th} moment of $m^0 \equiv m^0(n)$. Both (A1) and (A2) are less stringent and more natural than analogous conditions on $(2r-2)^{th}$ moments in the existing literature on AMP; see Remark 6.4, which also discusses (A3).

Given $\pi \in \mathcal{P}_1(r)$ from (A1) and $\tau_1 \in (0,\infty)$ from (A2), the state evolution parameters $(\tau_k^2 : k \in \mathbb{N})$ are defined inductively by

$$\tau_{k+1}^2 := \mathbb{E}\big(f_k(G_k,\bar{\gamma})^2\big), \tag{2.2}$$

where $G_k \sim N(0,\tau_k^2)$ and $\bar{\gamma} \sim \pi$ are independent. Since the functions f_k are Lipschitz and $\mathbb{E}(\bar{\gamma}^2)^{1/2} \leq \mathbb{E}(|\bar{\gamma}|^r)^{1/r} < \infty$ under (A1), it follows by induction that $\tau_k^2 \in [0,\infty)$ for all k.

We will make two further mild regularity assumptions. Suppose that if $r > 2$, then

(A4) $\pi\big(\{y \in \mathbb{R} : x \mapsto f_k(x,y) \text{ is non-constant}\}\big) > 0$ for each $k \in \mathbb{N}$.

This is a 'non-degeneracy' condition that ensures that $\tau_k^2 > 0$ for all $k \in \mathbb{N}$; see also Lemma 2.2 below. Henceforth, we will write $\mu \otimes \mu'$ for the product of two measures μ, μ'.

(A5) For each $k \in \mathbb{N}$, the set D_k of discontinuities of f_k' satisfies $(\lambda \otimes \pi)(D_k) = 0$, where λ denotes Lebesgue measure on $\mathbb{R}$.

This guarantees the existence of a deterministic limit for $b_k \equiv b_k(n)$ in (2.1) as $n \to \infty$ for each k (see Remark 2.4 below), and is satisfied by the functions f_k that are typically used in statistical applications, such as those based on soft-thresholding functions $\mathrm{ST}_t \colon u \mapsto \operatorname{sgn}(u)(|u|-t)_+$ for $t > 0$. See Section 6.1 for some technical remarks on (A1)–(A5), which can be skipped on a first reading.

We are now ready to state our first master theorem, which is a substantial result in random matrix theory. As mentioned in (i) above, this reveals in particular that the asymptotic distributional behaviour of the AMP iterates is governed by the scalar recursion (2.2).

Theorem 2.1. *Suppose that (A0)–(A5) hold for a sequence of symmetric AMP recursions (2.1) indexed by $n \in \mathbb{N}$. Then for each $k \in \mathbb{N}$, we have $d_r\big(\nu_n(h^k,\gamma), N(0,\tau_k^2) \otimes \pi\big) \xrightarrow{c} 0$ as $n \to \infty$, or equivalently*

$$\tilde{d}_r\big(\nu_n(h^k,\gamma), N(0,\tau_k^2)\otimes\pi\big) = \sup_{\psi \in \mathrm{PL}_2(r,1)} \left|\frac{1}{n}\sum_{i=1}^n \psi(h_i^k,\gamma_i) - \mathbb{E}\big(\psi(G_k,\bar{\gamma})\big)\right| \xrightarrow{c} 0 \quad \text{as } n \to \infty, \tag{2.3}$$

where $G_k \sim N(0,\tau_k^2)$ and $\bar{\gamma} \sim \pi$ are independent.

In the AMP literature, this conclusion is usually stated as

$$\frac{1}{n}\sum_{i=1}^n \psi(h_i^k,\gamma_i) \xrightarrow{a.s.} \mathbb{E}\big(\psi(G_k,\bar{\gamma})\big) \text{ as } n \to \infty, \text{ for every } \psi \in \mathrm{PL}_2(r). \tag{2.4}$$

In fact, $\xrightarrow{a.s.}$ can be strengthened to $\xrightarrow{c}$, and the resulting version of (2.4) is equivalent to (2.3); in other words, it can be upgraded automatically to a convergence statement that holds *uniformly* over the class $\mathrm{PL}_2(r,1)$ of pseudo-Lipschitz test functions. See Remarks 6.1 and 6.2 for further details.

To gain some insight into the form of the recursion (2.1) and its asymptotic characterisation in Theorem 2.1, suppose for simplicity that $\gamma \equiv \gamma(n) = 0 \in \mathbb{R}^n$ for all n, and first consider $k = 1$. Since $m^0 \equiv m^0(n)$ is independent of $W \equiv W(n)$ for each n by (A0), it follows that $h^1 \equiv h^1(n) = Wm^0$ is conditionally Gaussian given m^0. In fact, conditional on m^0,

$$h^1 \text{ and } h^{1,0} := \|m^0\|_n \tilde{Z} + \tilde{\zeta} m^0 = \tau_1 \tilde{Z} + \Delta^1 \text{ are identically distributed for each } n,$$

where $\tilde{Z} \sim N_n(0, I_n)$ is independent of $\tilde{\zeta} \sim N(0, 1/n)$, and where $\Delta^1 := (\|m^0\|_n - \tau_1)\tilde{Z} + \tilde{\zeta} m^0$; see Lemma 6.14, (6.5) and (6.20).

By (A2), $\|m^0\|_n \xrightarrow{c} \tau_1$ and $\|m^0\|_{n,r} = O_c(1)$ as $n \to \infty$, from which it follows (by the triangle inequality for $\|\cdot\|_{n,r}$) that $\|\Delta^1\|_{n,r} \equiv (n^{-1}\sum_{i=1}^n |\Delta_i^1|^r)^{1/r} \xrightarrow{c} 0$; see $\mathcal{H}_1(a)$ at the start of Section 6.5. This means that Δ^1 has asymptotically vanishing influence on the empirical distribution of the entries of $h^{1,0} \overset{d}{=} h^1 \in \mathbb{R}^n$ as $n \to \infty$, while the empirical distribution of the entries of $\tau_1 \tilde{Z}$ converges completely in d_r to $N(0, \tau_1^2)$ (essentially by the strong law of large numbers, or the concentration inequality in Lemma 7.12). This yields the conclusion of Theorem 2.1 for h^1, and also implies that

$$\|m^1\|_n^2 = \|f_1(h^1, 0)\|_n^2 = \frac{1}{n}\sum_{i=1}^n f_1(h_i^1, 0)^2 \xrightarrow{c} \mathbb{E}\big(f_1(G_1, 0)^2\big) = \tau_2^2,$$

$$\|m^1\|_{n,r}^r = \|f_1(h^1, 0)\|_{n,r}^r = \frac{1}{n}\sum_{i=1}^n |f_1(h_i^1, 0)|^r \xrightarrow{c} \mathbb{E}\big(|f_1(G_1, 0)|^r\big) < \infty.$$

These limits follow from the state evolution recursion (2.2) and the fact that f_1 is Lipschitz, whence $f_1^2, |f_1|^r \in \mathrm{PL}_2(r)$; see Corollary 7.21(b). Continuing inductively in this vein, we conclude that for each fixed $k \in \mathbb{N}$, the Gaussian distribution $N(0, \tau_k^2)$ in Theorem 2.1 is the d_r limit of the empirical distribution of the entries of $\breve{h}^k \equiv \breve{h}^k(n) \in \mathbb{R}^n$ in the 'toy' recursion

$$\breve{h}^1 := \breve{W}^0 m^0, \qquad \breve{m}^k := f_k(\breve{h}^k, \gamma), \qquad \breve{h}^{k+1} := \breve{W}^k \breve{m}^k \qquad \text{for } k \in \mathbb{N}, \tag{2.5}$$

where each $\breve{W}^k \equiv \breve{W}^k(n) \sim \mathrm{GOE}(n)$ is independent of m^0, γ ($= 0$ here) and $\breve{W}^0, \dots, \breve{W}^{k-1}$, and hence of $\breve{m}^k$.

On the other hand, observe that in the original recursion (2.1), the *same* $\mathrm{GOE}(n)$ matrix $W \equiv W(n)$ appears in every iteration, so W and m^k are not in general independent for $k \in \mathbb{N}$, and in fact Wm^k is not asymptotically Gaussian in the above sense. To compensate for this, the Onsager correction $-b_k m^{k-1}$ is designed specifically as a debiasing term to ensure that $h^{k+1} = Wm^k - b_k m^{k-1}$ has the same limiting behaviour as $\breve{h}^{k+1}$ in (2.5) above. Indeed, an important technical step in the proof of Theorem 2.1 is to characterise the conditional distribution of Wm^k given m^0, γ and the previous iterates $h^1, \dots, h^k$ (Proposition 6.11), and then show that the 'non-Gaussian components' thereof are asymptotically cancelled out by the Onsager term.

This ingenious conditioning technique was first developed by Bolthausen (2014) and Bayati and Montanari (2011), and later used extensively in the analysis of various other AMP iterations in which W is drawn from a rotationally invariant matrix ensemble. For example, Berthier et al. (2020) introduced a 'Long AMP' recursion in which each iterate h^{k+1} is defined more explicitly in terms of the Gaussian part of the conditional distribution of $Wf_k(h^k, \gamma)$. For the symmetric AMP recursion (2.1), the relevant results on conditional distributions are stated in Section 6.2, where we discuss the subtleties in their derivation, and then rigorously proved in Section 6.3.

We give a technical summary of the proof of Theorem 2.1 in Section 6.4, where the key result is Proposition 6.16, and defer the formal arguments to Section 6.5. The proof proceeds by induction on $k \in \mathbb{N}$ and actually establishes a more general result (Theorem 2.3 below) that implies Theorem 2.1: in particular, for fixed $k \in \mathbb{N}$, the joint empirical distribution of the components of $h^1, \dots, h^k \in \mathbb{R}^n$ converges completely in d_r to a Gaussian limit $N_k(0, \bar{\mathrm{T}}^{[k]})$ as $n \to \infty$.

The sequence $(\bar{\mathrm{T}}^{[k]} \in \mathbb{R}^{k\times k} : k \in \mathbb{N})$ of covariance matrices is defined recursively as an extension of the state evolution (2.2), and we will see from (2.9) below that the covariances can be characterised as limits of inner products $\langle m^{k-1}, m^{\ell-1}\rangle_n$ between iterates $m^{k-1}, m^{\ell-1}$ in (2.1) as $n \to \infty$. First, let $G_1 \sim N(0, \tau_1^2)$ and $\bar{\mathrm{T}}_{1,1} := \tau_1^2$, so that $\bar{\mathrm{T}}^{[1]} \equiv \mathrm{Var}(G_1) = \bar{\mathrm{T}}_{1,1} \geq 0$. For a general $k \geq 2$, suppose inductively that we have already defined a non-negative definite $\bar{\mathrm{T}}^{[k-1]} \in \mathbb{R}^{(k-1)\times(k-1)}$ with entries

$\bar{\mathrm{T}}^{[k-1]}_{ij} = \bar{\mathrm{T}}_{i,j}$ for $1 \leq i,j \leq k-1$, and then let

$$\bar{\mathrm{T}}_{k,\ell} = \bar{\mathrm{T}}_{\ell,k} := \begin{cases} \mathbb{E}\bigl(F_0(\bar{\gamma}) \cdot f_{k-1}(G_{k-1},\bar{\gamma})\bigr) & \text{for } \ell = 1 \\ \mathbb{E}\bigl(f_{\ell-1}(G_{\ell-1},\bar{\gamma}) \cdot f_{k-1}(G_{k-1},\bar{\gamma})\bigr) & \text{for } \ell = 2,\dots,k, \end{cases} \tag{2.6}$$

where F_0 is as in (A3) and $\bar{\gamma} \sim \pi$ is independent of $(G_1,\dots,G_{k-1}) \sim N_{k-1}(0,\bar{\mathrm{T}}^{[k-1]})$. Define $\bar{\mathrm{T}}^{[k]}$ to be the $k \times k$ matrix with entries $\bar{\mathrm{T}}^{[k]}_{ij} = \bar{\mathrm{T}}_{i,j}$ for $1 \leq i,j \leq k$, so that $\bar{\mathrm{T}}^{[k-1]}$ is the top-left principal $(k-1) \times (k-1)$ submatrix of $\bar{\mathrm{T}}^{[k]}$. For every $a \equiv (a_1,\dots,a_k) \in \mathbb{R}^k$, we have

$$a^\top \bar{\mathrm{T}}^{[k]} a = \mathbb{E}\Bigl\{\bigl(a_1 F_0(\bar{\gamma}) + \textstyle\sum_{\ell=2}^k a_\ell f_{\ell-1}(G_{\ell-1},\bar{\gamma})\bigr)^2\Bigr\} + a_1^2\bigl\{\tau_1^2 - \mathbb{E}\bigl(F_0(\bar{\gamma})^2\bigr)\bigr\} \geq 0 \tag{2.7}$$

since $\mathbb{E}\bigl(F_0(\bar{\gamma})^2\bigr) \leq \tau_1^2$ by (A3), so $\bar{\mathrm{T}}^{[k]} \in \mathbb{R}^{k\times k}$ is non-negative definite. In fact, we have the following:

Lemma 2.2. *Under (A4), $\bar{\mathrm{T}}^{[k]} \in \mathbb{R}^{k\times k}$ is positive definite and hence invertible for every $k \in \mathbb{N}$.*

The proof of this fact is given in Section 6.6. By induction, we have $\tau_k^2 = \mathbb{E}\bigl(f_{k-1}(G_{k-1},\bar{\gamma})^2\bigr) = \bar{\mathrm{T}}_{k,k} > 0$ for all $k \in \mathbb{N}$, so (2.6) does indeed extend (2.2). Our second master theorem is the following:

Theorem 2.3. *Under the hypotheses of Theorem 2.1, $d_r\bigl(\nu_n(h^1,\dots,h^k,\gamma), N_k(0,\bar{\mathrm{T}}^{[k]}) \otimes \pi\bigr) \xrightarrow{c} 0$ for each fixed $k \in \mathbb{N}$ as $n \to \infty$, or equivalently*

$$\widetilde{d}_r\bigl(\nu_n(h^1,\dots,h^k,\gamma), N_k(0,\bar{\mathrm{T}}^{[k]}) \otimes \pi\bigr) = \sup_{\psi \in \mathrm{PL}_{k+1}(r,1)} \left|\frac{1}{n}\sum_{i=1}^n \psi(h_i^1,\dots,h_i^k,\gamma_i) - \mathbb{E}\bigl(\psi(G_1,\dots,G_k,\bar{\gamma})\bigr)\right| \xrightarrow{c} 0 \tag{2.8}$$

as $n \to \infty$, where $(G_1,\dots,G_k) \sim N_k(0,\bar{\mathrm{T}}^{[k]})$ and $\bar{\gamma} \sim \pi$ are independent. In particular,

$$\bar{\mathrm{T}}_{k,\ell} = \operatorname*{c\text{-}lim}_{n\to\infty} \langle m^{k-1}, m^{\ell-1}\rangle_n \quad \textit{for all } k,\ell \geq 1. \tag{2.9}$$

In statistical applications featuring AMP iterates of the form $\bigl(v^k \equiv v^k(n) : k,n \in \mathbb{N}\bigr)$, we sometimes require joint convergence guarantees of the form (2.8) in addition to results along the lines of the original Theorem 2.1. For example, see Step (II) of the analysis of (3.28) in Section 3.3 and also Step 3 of the general recipe of Section 4.4. In both cases, we need to track the limiting covariances as well as the limiting variances (state evolution parameters) to show that $\lim_{k\to\infty}\lim_{n\to\infty} \|v^{k+1} - v^k\|_n \xrightarrow{c} 0$, i.e. that the asymptotic differences between successive AMP iterates become negligible for large k.

Remark 2.4. The precise form of the Onsager coefficient b_k in (2.1) is essentially due to Stein's lemma; see (6.19) and Proposition 6.16(g) below. The latter shows that under (A5),

$$b_k(n) = \langle f_k'(h^k,\gamma)\rangle_n \xrightarrow{c} \mathbb{E}\bigl(f_k'(G_k,\bar{\gamma})\bigr) =: \bar{b}_k$$

for each k as $n \to \infty$. The conclusions of Theorems 2.1 and 2.3 remain valid if we replace $b_k \equiv b_k(n)$ with $\bar{b}_k$ in the recursion (2.1) for all k,n, in which case (A5) is no longer needed.

For $1 \leq j,\ell \leq k$, since $\psi\colon (x_1,\dots,x_k,y) \mapsto x_j x_\ell$ lies in $\mathrm{PL}_{k+1}(2) \subseteq \mathrm{PL}_{k+1}(r)$, (2.8) implies that $\langle h^j, h^\ell\rangle_n \xrightarrow{c} \mathbb{E}(G_j G_\ell) = \bar{\mathrm{T}}_{j,\ell}$. Thus, the limiting covariance structure of $h^1,\dots,h^k$ is given by $\mathrm{T}^{[k]}$, which in general is not a diagonal matrix. By contrast, while $\breve{h}^k$ in the toy recursion (2.5) has the same asymptotics as h^k as $n \to \infty$, it turns out that $\breve{h}^1,\dots,\breve{h}^k$ are asymptotically independent, in the sense that the d_r limit of the joint empirical distribution of their components is a centred Gaussian with covariance $\mathrm{diag}(\tau_1^2,\dots,\tau_k^2)$.

2.2 Asymmetric AMP

For $n,p \in \mathbb{N}$, the abstract asymmetric AMP recursion (2.10) below is based on a matrix $W \in \mathbb{R}^{n\times p}$, two vectors $\beta \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}^n$ of auxiliary information and two sequences $(g_k, f_{k+1} : k \in \mathbb{N}_0)$ of

Lipschitz functions $g_k, f_{k+1}\colon \mathbb{R}^2 \to \mathbb{R}$. Given $q^{-1} := 0 \in \mathbb{R}^n$, $b_0 \in \mathbb{R}$ and $m^0 \in \mathbb{R}^p$, we inductively define

$$\begin{aligned} e^k &:= Wm^k - b_k q^{k-1}, & q^k &:= g_k(e^k, \gamma), & c_k &:= n^{-1} \textstyle\sum_{i=1}^n g_k'(e_i^k, \gamma_i), \\ h^{k+1} &:= W^\top q^k - c_k m^k, & m^{k+1} &:= f_{k+1}(h^{k+1}, \beta), & b_{k+1} &:= n^{-1} \textstyle\sum_{j=1}^p f_{k+1}'(h_j^{k+1}, \beta_j) \end{aligned} \tag{2.10}$$

for $k \in \mathbb{N}_0$. Here, $g_k', f_{k+1}'\colon \mathbb{R}^2 \to \mathbb{R}$ are bounded, Borel measurable functions that agree with the partial derivatives of g_k, f_{k+1} respectively with respect to their first arguments, wherever the latter are defined.

A master theorem for (2.10) is stated below as Theorem 2.5, whose hypotheses and conclusions are similar to those of Theorems 2.1 and 2.3 for the symmetric iteration (2.1). Consider a sequence of recursions (2.10) indexed by $n \in \mathbb{N}$ and $p \equiv p_n$, for which $n/p \to \delta \in (0, \infty)$ as $n \to \infty$. In this asymptotic regime, suppose that there exist $r \in [2, \infty)$ and $\sigma_0 \in (0, \infty)$ for which the following analogues of (A0)–(A5) hold:

- (B0) For each n, the matrix $W \equiv W(n)$ has entries $W_{ij} \overset{\mathrm{iid}}{\sim} N(0, 1/n)$ for $1 \le i \le n$ and $1 \le j \le p$, and is independent of $(m^0, \beta, \gamma) \equiv \big(m^0(n), \beta(n), \gamma(n)\big)$.
- (B1) There exist probability distributions $\pi_{\bar\beta}, \pi_{\bar\gamma} \in \mathcal{P}_1(r)$ such that writing $\nu_p(\beta)$ and $\nu_n(\gamma)$ for the empirical distributions of the components of $\beta \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}^n$ respectively, we have $d_r\big(\nu_p(\beta), \pi_{\bar\beta}\big) \xrightarrow{c} 0$ and $d_r\big(\nu_n(\gamma), \pi_{\bar\gamma}\big) \xrightarrow{c} 0$.
- (B2) $\sqrt{p/n}\,\|m^0\|_p \equiv (n^{-1} \sum_{j=1}^p |m_j^0|^2)^{1/2} \xrightarrow{c} \sigma_0$ and $\|m^0\|_{p,r} \equiv (p^{-1} \sum_{j=1}^p |m_j^0|^r)^{1/r} = O_c(1)$.
- (B3) There exists a Lipschitz $F_0\colon \mathbb{R} \to \mathbb{R}$ such that taking $\bar\beta \sim \pi_{\bar\beta}$, we have $\mathbb{E}\big(F_0(\bar\beta)^2\big) \le \sigma_0^2$ and $\langle m^0, \phi(\beta) \rangle_p = p^{-1} \sum_{j=1}^p f_0(h_j^0, \beta_j)\, \phi(\beta_j) \xrightarrow{c} \mathbb{E}\big(F_0(\bar\beta)\phi(\bar\beta)\big)$ for all Lipschitz $\phi\colon \mathbb{R} \to \mathbb{R}$.
- (B4) For each $k \in \mathbb{N}_0$, we have $\pi_{\bar\gamma}\big(\{y \in \mathbb{R} : x \mapsto g_k(x, y) \text{ is non-constant}\}\big) > 0$ and $\pi_{\bar\beta}\big(\{y \in \mathbb{R} : x \mapsto f_{k+1}(x, y) \text{ is non-constant}\}\big) > 0$.
- (B5) For each $k \in \mathbb{N}_0$, writing D_k, C_{k+1} for the sets of discontinuities of g_k', f_{k+1}' respectively, we have $(\lambda \otimes \pi_{\bar\gamma})(D_k) = (\lambda \otimes \pi_{\bar\beta})(C_{k+1}) = 0$, where λ denotes Lebesgue measure on $\mathbb{R}$.

State evolution: With $\sigma_0 > 0$ as above, inductively define

$$\tau_{k+1}^2 := \mathbb{E}\big(g_k(G_k^\sigma, \bar\gamma)^2\big) \quad \text{and} \quad \sigma_{k+1}^2 := \delta^{-1}\, \mathbb{E}\big(f_{k+1}(G_{k+1}^\tau, \bar\beta)^2\big) \tag{2.11}$$

for $k \in \mathbb{N}_0$, where we take $G_k^\sigma \sim N(0, \sigma_k^2)$ to be independent of $\bar\beta \sim \pi_{\bar\beta}$, and $G_{k+1}^\tau \sim N(0, \tau_{k+1}^2)$ to be independent of $\bar\gamma \sim \pi_{\bar\gamma}$.

Limiting covariance structure: Let $\bar\Sigma^{[1]} \equiv \bar\Sigma_{0,0} := \sigma_0^2$ and $\bar{\mathrm{T}}^{[1]} \equiv \bar{\mathrm{T}}_{1,1} := \tau_1^2$, and for a general $k \in \mathbb{N}$, suppose inductively that we have already defined non-negative definite matrices $\bar\Sigma^{[k]}, \bar{\mathrm{T}}^{[k]} \in \mathbb{R}^{k \times k}$ with entries $\bar\Sigma_{ij}^{[k]} = \bar\Sigma_{i-1,j-1}$ and $\bar{\mathrm{T}}_{ij}^{[k]} = \bar{\mathrm{T}}_{i,j}$ for $1 \le i, j \le k$. Then let

$$\bar\Sigma_{k,\ell} = \bar\Sigma_{\ell,k} := \begin{cases} \delta^{-1}\, \mathbb{E}\big(F_0(\bar\beta) \cdot f_k(G_k^\tau, \bar\beta)\big) & \text{for } \ell = 0 \\ \delta^{-1}\, \mathbb{E}\big(f_\ell(G_\ell^\tau, \bar\beta) \cdot f_k(G_k^\tau, \bar\beta)\big) & \text{for } \ell = 1, \dots, k, \end{cases} \tag{2.12}$$

where $(G_1^\tau, \dots, G_k^\tau) \sim N_k(0, \bar{\mathrm{T}}^{[k]})$ is independent of $\bar\gamma \sim \pi_{\bar\gamma}$, and define $\bar\Sigma^{[k+1]} \in \mathbb{R}^{(k+1)\times(k+1)}$ by $\bar\Sigma_{ij}^{[k+1]} := \bar\Sigma_{i-1,j-1}$ for $1 \le i, j \le k+1$. As in (2.7), it is easily verified that $\bar\Sigma^{[k+1]}$ is non-negative definite. In addition, let

$$\bar{\mathrm{T}}_{k+1,\ell} = \bar{\mathrm{T}}_{\ell,k+1} := \mathbb{E}\big(g_{\ell-1}(G_{\ell-1}^\sigma, \bar\gamma) \cdot g_k(G_k^\sigma, \bar\gamma)\big) \qquad \text{for } \ell = 1, \dots, k+1, \tag{2.13}$$

where $(G_0^\sigma, \dots, G_k^\sigma) \sim N_{k+1}(0, \bar\Sigma^{[k+1]})$ is independent of $\bar\beta \sim \pi_{\bar\beta}$, and define $\bar{\mathrm{T}}_{ij}^{[k+1]} := \bar{\mathrm{T}}_{i,j}$ for $1 \le i, j \le k+1$, so that the resulting matrix $\bar{\mathrm{T}}^{[k+1]} \in \mathbb{R}^{(k+1)\times(k+1)}$ is again non-negative definite. Under (B4), it can be shown as in Lemma 2.2 that $\bar\Sigma^{[k]}, \bar{\mathrm{T}}^{[k]}$ are positive definite for all $k \in \mathbb{N}$, and also that (2.12)–(2.13) extends (2.11), with $\sigma_{k-1}^2 = \bar\Sigma_{k-1,k-1} > 0$ and $\tau_k^2 = \bar{\mathrm{T}}_{k,k} > 0$ for all k.

Theorem 2.5. *Suppose that (B0)–(B5) hold for a sequence of asymmetric AMP recursions (2.10) indexed by $n \in \mathbb{N}$ and $p \equiv p_n$ with $n/p \to \delta \in (0, \infty)$. Then for each fixed $k \in \mathbb{N}_0$, we have*

$$\begin{aligned}\widetilde{d}_r\big(\nu_n(e^k, \gamma), N(0, \sigma_k^2) \otimes \pi_{\bar{\gamma}}\big) &= \sup_{\psi \in \mathrm{PL}_2(r,1)} \left|\frac{1}{n}\sum_{i=1}^n \psi(e_i^k, \gamma_i) - \mathbb{E}\big(\psi(G_k^\sigma, \bar{\gamma})\big)\right| \xrightarrow{c} 0,\\ \widetilde{d}_r\big(\nu_p(h^{k+1}, \beta), N(0, \tau_{k+1}^2) \otimes \pi_{\bar{\beta}}\big) &= \sup_{\psi \in \mathrm{PL}_2(r,1)} \left|\frac{1}{p}\sum_{j=1}^p \psi(h_j^{k+1}, \beta_j) - \mathbb{E}\big(\psi(G_{k+1}^\tau, \bar{\beta})\big)\right| \xrightarrow{c} 0,\end{aligned} \tag{2.14}$$

$$\begin{aligned}&\widetilde{d}_r\big(\nu_n(e^0, \dots, e^k, \beta), N_{k+1}(0, \bar{\Sigma}^{[k+1]}) \otimes \pi_{\bar{\beta}}\big)\\ &\qquad= \sup_{\psi \in \mathrm{PL}_{k+2}(r,1)} \left|\frac{1}{n}\sum_{i=1}^n \psi(e_i^0, \dots, e_i^k, \gamma_i) - \mathbb{E}\big(\psi(G_0^\sigma, \dots, G_k^\sigma, \bar{\gamma})\big)\right| \xrightarrow{c} 0,\\ &\widetilde{d}_r\big(\nu_p(h^1, \dots, h^{k+1}, \beta), N_{k+1}(0, \bar{\mathrm{T}}^{[k+1]}) \otimes \pi_{\bar{\beta}}\big)\\ &\qquad= \sup_{\psi \in \mathrm{PL}_{k+2}(r,1)} \left|\frac{1}{p}\sum_{j=1}^p \psi(h_j^1, \dots, h_j^{k+1}, \beta_j) - \mathbb{E}\big(\psi(G_1^\tau, \dots, G_{k+1}^\tau, \bar{\beta})\big)\right| \xrightarrow{c} 0\end{aligned} \tag{2.15}$$

as $n \to \infty$. Equivalent statements hold with d_r in place of $\widetilde{d}_r$.

Together with the master theorems in Section 2.1, Theorem 2.5 can be generalised to abstract AMP recursions with matrix-valued iterates; see Section 6.7.

Similarly to the discussion after Theorem 2.1, one can argue that for each $k \in \mathbb{N}_0$, the Gaussian distributions $N(0, \sigma_k^2)$ and $N(0, \tau_{k+1}^2)$ in (2.14) are the d_r limits of the empirical distributions of the entries of $\breve{e}^k \in \mathbb{R}^n$ and $\breve{h}^{k+1} \in \mathbb{R}^p$ respectively in the toy recursion

$$\breve{e}^0 := \breve{W}^0 m^0, \qquad \breve{h}^{k+1} := \breve{W}^k g_k(\breve{e}^k, \gamma), \qquad \breve{e}^{k+1} := \breve{W}^{k+1} f_k(\breve{h}^{k+1}, \beta) \qquad \text{for } k \in \mathbb{N}_0 \tag{2.16}$$

as $n, p \to \infty$ with $n/p \to \delta$. Here, each iteration features a new matrix with i.i.d. $N(0, 1/n)$ entries that is independent of everything thus far. In the original abstract iteration (2.10), where the same Gaussian matrix W is used throughout, the Onsager correction terms $-b_k q^{k-1}$ and $-c_k m^k$ are designed to ensure that $e^k \in \mathbb{R}^n$ and $h^{k+1} \in \mathbb{R}^p$ have the same limiting behaviour as $\breve{e}^k$ and $\breve{h}^{k+1}$ respectively. We note however that the asymptotic joint empirical distributions in (2.15) are in general different from those in (2.16). Indeed, the limiting covariance matrices in (2.16) are diagonal whereas $\bar{\Sigma}$ and $\bar{\mathrm{T}}$ in (2.12)–(2.13) are usually not diagonal; see the end of Section 2.1 for a similar comparison of the symmetric recursions (2.1) and (2.5).

One way to establish Theorem 2.5 is to analyse the asymmetric recursion (2.10) directly, by adapting the techniques and arguments from the proof of Theorem 2.3 for the symmetric iteration (2.1). An important first step is to obtain an analogue of Proposition 6.11 that characterises the conditional distribution of each of the iterates in (2.10), given the inputs m^0, β, γ and all the previous iterates. This then sets up an inductive proof along the lines of Proposition 6.16 (Bayati and Montanari, 2011). Rush and Venkataramanan (2018) established a finite-sample version of Theorem 2.5 under finite-sample analogues of its hypotheses (see Remark 6.3).

There is an alternative derivation of Theorem 2.5 that proceeds by first embedding (2.10) within a suitable symmetric recursion (featuring a $\mathrm{GOE}(n+p)$ matrix), whose output at iteration $k \in \mathbb{N}_0$ contains h^ℓ when $k = 2\ell$ and e^ℓ when $k = 2\ell + 1$ (Javanmard and Montanari, 2013; Berthier et al., 2020). The construction of this augmented recursion is based on a slightly more general version of the original symmetric iteration (2.1) that offers the additional flexibility to apply (two) different Lipschitz functions to different components of each AMP iterate.

3 Low-rank matrix estimation

3.1 An AMP algorithm for estimating a symmetric rank-one matrix

In this subsection, we will motivate and analyse an AMP algorithm for reconstructing a symmetric rank-one matrix based on an observation

$$A \equiv A(n) = \frac{\lambda}{n} vv^\top + W \in \mathbb{R}^{n\times n} \tag{3.1}$$

for some $n \in \mathbb{N}$, where $\lambda > 0$ is a deterministic scalar, $v \equiv v(n) \in \mathbb{R}^n$ is the signal (or 'spike') that we wish to estimate, and $W \equiv W(n) \sim \mathrm{GOE}(n)$ is a noise matrix. The asymptotic setting of interest to us here is one where $\|v\|_n \equiv n^{-1/2}\,\|v\|$ converges to 1 as $n \to \infty$; see (3.4) below.

A natural estimator of v is a principal eigenvector $\hat{\varphi} \equiv \varphi^1(A) \in \mathbb{R}^n$ (with $\|\hat{\varphi}\|_n = 1$) corresponding to the largest eigenvalue $\lambda_1(A)$ of the observation matrix A. A cornerstone of the spectral theory of such 'deformed' GOE matrices is the so-called 'BBP' phase transition. This was first established in the seminal paper of Baik et al. (2005) and later explored in greater generality by Baik and Silverstein (2006), Féral and Péché (2007), Capitaine et al. (2009) and Benaych-Georges and Nadakuditi (2011), among many others. See Johnstone and Paul (2018) for an accessible summary of this line of work, which reveals that in the limiting regime where $\|v\|_n$ converges to 1, the eigenstructure of $A \equiv A(n)$ for large n exhibits two different types of qualitative behaviour depending on whether $\lambda \le 1$ or $\lambda > 1$. In particular, when $n \to \infty$, it follows from the concentration results in Knowles and Yin (2013, Theorems 2.7 and 6.3) that

$$\lambda_1(A) \xrightarrow{c} \begin{cases} \lambda + \lambda^{-1} > 2 & \text{if } \lambda > 1 \\ 2 & \text{if } \lambda \in (0,1], \end{cases} \qquad \frac{|\langle \hat{\varphi}, v\rangle|}{\|\hat{\varphi}\|\,\|v\|} \xrightarrow{c} \begin{cases} \sqrt{1-\lambda^{-2}} & \text{if } \lambda > 1 \\ 0 & \text{if } \lambda \in (0,1]; \end{cases} \tag{3.2}$$

see also Peng (2012, Theorem 3.1) for the former and Corollary 3.4 below for the latter.

In the 'supercritical' phase when $\lambda > 1$, the effect of the spike v can be seen in the limiting expressions above: with high probability, $\hat{\varphi}$ is at least partially aligned with v (although it does not estimate v consistently) and $\lambda_1(A)$ is an outlier that is separated from the 'bulk' of the spectrum of A. Indeed, the remaining eigenvalues of A are asymptotically distributed according to the Wigner semicircle law on $[-2, 2]$, and it can be shown that the second-largest eigenvalue $\lambda_2(A)$ of $A \equiv A(n)$ satisfies $\lambda_2(A) \xrightarrow{c} 2$ as $n \to \infty$, so the limiting spectral gap $\lambda_1(A) - \lambda_2(A)$ is strictly positive.

On the other hand, in the 'subcritical' phase when $\lambda \le 1$, the noise matrix W obscures the signal in (3.1) to such an extent that $\hat{\varphi}$ is asymptotically uninformative as an estimator of v, as evidenced by the asymptotic orthogonality in (3.2), and $\lambda_1(A)$ remains attached to the bulk of the eigenvalues of A. In this low signal-to-noise regime, the limits for $\lambda_1(A)$ and $\hat{\varphi}$ in (3.2) are the same as for the leading eigenvalue and eigenvector of W respectively.

A further limitation of the classical spectral estimator $\hat{\varphi}$ is that it is unable to exploit any additional information about the structure of v that may be relevant for inference. For example, in some matrix estimation problems such as hidden clique detection and non-negative or sparse principal component analysis, there are natural constraints that force v to be non-negative or sparse, or to lie in some finite set such as $\{0,1\}^n$ (Alon et al., 1998; Zou et al., 2006; Vu and Lei, 2013; Deshpande and Montanari, 2015; Montanari and Richard, 2016). A Bayesian approach to modelling a structured signal v is to assume that its components are drawn from some suitable prior distribution that is fully or partially known. However, for general priors, a practical issue is the lack of efficient (i.e. polynomial-time) algorithms for computing or accurately approximating the Bayes estimator of v with respect to quadratic loss, namely the posterior mean $\mathbb{E}(v \mid A)$.

We will now present a generic (and computationally feasible) AMP procedure (3.3) for estimating v (cf. Deshpande and Montanari, 2014; Deshpande et al., 2016; Montanari and Venkataramanan, 2021), and obtain an exact characterisation of its asymptotic performance in terms of a state evolution recursion (Theorem 3.1 and Corollary 3.2). Guided by these theoretical guarantees, we will explain in Sections 3.2 and 3.3 how the inputs to the algorithm can be specialised further to take advantage of different types of prior information, and thereby produce estimators that outperform $\hat{\varphi}$ in terms of asymptotic mean squared error.

Let $(g_k)_{k=0}^{\infty}$ be a sequence of Lipschitz functions on $\mathbb{R}$ with corresponding weak derivatives g_k'. Given $\hat{v}^{-1} \equiv \hat{v}^{-1}(n) := 0 \in \mathbb{R}^n$ and an initialiser $v^0 \equiv v^0(n) \in \mathbb{R}^n$ for some $n \in \mathbb{N}$, we recursively define $v^k \equiv v^k(n) \in \mathbb{R}^n$, $b_k \equiv b_k(n) \in \mathbb{R}$ and $\hat{v}^{k+1} \equiv \hat{v}^{k+1}(n) \in \mathbb{R}^n$ by

$$\hat{v}^k := g_k(v^k), \qquad b_k := \langle g_k'(v^k)\rangle_n = \frac{1}{n}\sum_{i=1}^n g_k'(v_i^k), \qquad v^{k+1} := A\hat{v}^k - b_k\hat{v}^{k-1} \tag{3.3}$$

for $k \in \mathbb{N}_0$. This has a very similar form to the abstract recursion (2.1) that we studied in Section 2.1, the main difference being that (3.3) is a valid algorithm with the data matrix $A \equiv A(n)$ in place of the unobserved noise matrix $W \equiv W(n)$.

As mentioned in the Introduction, we can view (3.3) as a generalised power iteration, in which the additional Onsager correction term $-b_k\hat{v}^{k-1}$ is crucial for ensuring that the iterates v^k have the desired asymptotic distributional properties. In fact, we will see in Section 3.3 that for a specific choice of linear functions g_k given by (3.27), the corresponding recursion (3.3) is asymptotically equivalent to a standard power iteration that converges to the principal eigenvector $\hat{\varphi}$ of A.

To set up our asymptotic framework, consider a sequence of recursions (3.3) indexed by $n \in \mathbb{N}$, for which the following conditions hold:

(M0) The noise matrix $W \equiv W(n) \sim \mathrm{GOE}(n)$ in (3.1) is independent of $(\hat{v}^0, v) \equiv \bigl(\hat{v}^0(n), v(n)\bigr)$ for each n.

(M1) There exist $\mu_0, \sigma_0 \in \mathbb{R}$ and independent random variables U, V with $\mathbb{E}(U^2) = \mathbb{E}(V^2) = 1$, such that

$$\sup_{\psi \in \mathrm{PL}_2(2,1)} \left|\frac{1}{n}\sum_{i=1}^n \psi(v_i^0, v_i) - \mathbb{E}\bigl\{\psi\bigl(\mu_0 V + \sigma_0 U, V\bigr)\bigr\}\right| \xrightarrow{c} 0.$$

In other words, writing $\bar{\mu}^0$ for the distribution of $(\mu_0 V + \sigma_0 U, V)$, and $\nu_n(v^0, v)$ for the joint empirical distribution of the components of $v^0, v \in \mathbb{R}^n$ for $n \in \mathbb{N}$, we have

$$\tilde{d}_2\bigl(\nu_n(v^0, v), \bar{\mu}^0\bigr) \xrightarrow{c} 0 \quad \text{or equivalently} \quad d_2\bigl(\nu_n(v^0, v), \bar{\mu}^0\bigr) \xrightarrow{c} 0.$$

(M2) For each $k \in \mathbb{N}$, the function $g_k' \colon \mathbb{R} \to \mathbb{R}$ is continuous Lebesgue almost everywhere, i.e. the set of discontinuities of g_k' has Lebesgue measure 0.

Henceforth, we will write π for the distribution of V, which can be viewed as the 'limiting prior distribution' of the components of the signal $v \equiv v(n)$. Note that while v is only identifiable up to sign in the original spiked model (3.1), knowledge of π may help us to distinguish v from $-v$ in the limit $n \to \infty$, for example if π has non-zero mean. By considering the $\mathrm{PL}_2(2)$ functions $(x, y) \mapsto y^2$, $(x, y) \mapsto xy$ and $(x, y) \mapsto x^2$, we deduce from (M1) that

$$\|v\|_n^2 = \frac{1}{n}\sum_{i=1}^n v_i^2 \xrightarrow{c} \mathbb{E}(V^2) = 1, \tag{3.4}$$

$$\lambda\langle\hat{v}^0, v\rangle_n \xrightarrow{c} \lambda\mathbb{E}\bigl(Vg_0(\mu_0 V + \sigma_0 U)\bigr) =: \mu_1 \quad \text{and} \quad \|\hat{v}^0\|_n^2 \xrightarrow{c} \mathbb{E}\bigl(g_0(\mu_0 V + \sigma_0 U)^2\bigr) =: \sigma_1^2. \tag{3.5}$$

State evolution: Starting with $\mu_1 \in \mathbb{R}$ and $\sigma_1 \in [0, \infty)$, we inductively define state evolution parameters $\mu_k \in \mathbb{R}$ and $\sigma_k \in [0, \infty)$ for $k \in \mathbb{N}$ by

$$\mu_{k+1} := \lambda\mathbb{E}\bigl(Vg_k(\mu_k V + \sigma_k G)\bigr), \qquad \sigma_{k+1}^2 := \mathbb{E}\bigl(g_k(\mu_k V + \sigma_k G)^2\bigr), \tag{3.6}$$

where $V \sim \pi$ and $G \sim N(0,1)$ are independent. Note that since each g_k is Lipschitz and $\mathbb{E}(V^2) = \mathbb{E}(G^2) = 1$, we indeed have $\mu_k \in \mathbb{R}$ and $\sigma_k \in [0,\infty)$ for all k by induction; we will see below that these represent the *effective signal strength* and *effective noise level* respectively at iteration k.

Limiting covariance structure: We now extend (3.6) by specifying the covariance matrices of the limiting Gaussian distributions in Theorem 3.1 below. Let $\bar{\Sigma}^{[1]} = \bar{\Sigma}_{1,1} := \sigma_1^2 \geq 0$. For a general $k \geq 2$, suppose inductively that we have already defined a non-negative definite $\bar{\Sigma}^{[k-1]} \in \mathbb{R}^{(k-1)\times(k-1)}$ with entries $\bar{\Sigma}^{[k-1]}_{ij} = \bar{\Sigma}_{i,j}$ for $1 \leq i,j \leq k-1$, and then let

$$\bar{\Sigma}_{k,\ell} = \bar{\Sigma}_{\ell,k} := \begin{cases} \mathbb{E}\bigl(g_0(\mu_0 V + \sigma_0 U)\cdot g_{k-1}(\mu_{k-1}V + \sigma_{k-1}G_{k-1})\bigr) & \text{for } \ell = 1 \\ \mathbb{E}\bigl(g_{\ell-1}(\mu_{\ell-1}V + \sigma_{\ell-1}G_{\ell-1})\cdot g_{k-1}(\mu_{k-1}V + \sigma_{k-1}G_{k-1})\bigr) & \text{for } \ell = 2,\dots,k, \end{cases} \tag{3.7}$$

where $(\sigma_1 G_1,\dots,\sigma_{k-1}G_{k-1}) \sim N_{k-1}(0,\bar{\Sigma}^{[k-1]})$ is independent of (U,V) from (M1). Let $\bar{\Sigma}^{[k]}$ be the $k \times k$ matrix with entries $\bar{\Sigma}^{[k]}_{ij} := \bar{\Sigma}_{i,j}$ for $1 \leq i,j \leq k$, so that $\bar{\Sigma}^{[k-1]}$ is the top-left principal $(k-1)\times(k-1)$ submatrix of $\bar{\Sigma}^{[k]}$. It can be verified as in (2.7) that $\bar{\Sigma}^{[k]}$ is non-negative definite. By induction, $\sigma_k^2 = \mathbb{E}\bigl(g_{k-1}(\mu_{k-1}V + \sigma_{k-1}G)^2\bigr) = \bar{\Sigma}_{k,k}$ for all $k \in \mathbb{N}$, so (3.7) does indeed extend (3.6).

We are now ready to state the main result of this subsection, which for each $k \in \mathbb{N}$ establishes the 2-Wasserstein (d_2) limit of the joint empirical distributions of the components of $v^0, v^1, \dots, v^k, v \in \mathbb{R}^n$ as $n \to \infty$.

Theorem 3.1. *Suppose that (M0)–(M2) hold for a sequence of AMP iterations (3.3), where for each $n \in \mathbb{N}$, the symmetric matrix $A \equiv A(n)$ is generated according to the spiked model (3.1) for some fixed $\lambda > 0$ that does not depend on n. Then for each $k \in \mathbb{N}$, we have*

$$\sup_{\psi \in \mathrm{PL}_{k+2}(2,1)} \left| \frac{1}{n}\sum_{i=1}^n \psi(v_i^0, v_i^1 \dots, v_i^k, v_i) - \mathbb{E}\bigl(\psi(\mu_0 V + \sigma_0 U, \mu_1 V + \sigma_1 G_1, \dots, \mu_k V + \sigma_k G_k, V)\bigr)\right| \xrightarrow{c} 0 \tag{3.8}$$

as $n \to \infty$, where $(\sigma_1 G_1,\dots,\sigma_k G_k) \sim N_k(0,\bar{\Sigma}^{[k]})$ is independent of (U,V) from (M1). In other words, writing $\breve{\nu}^k$ for the distribution of $(\mu_0 V + \sigma_0 U, \mu_1 V + \sigma_1 G_1, \dots, \mu_k V + \sigma_k G_k, V)$, we have

$$\widetilde{d}_2\bigl(\nu_n(v^0, v^1, \dots, v^k, v), \breve{\nu}^k\bigr) \xrightarrow{c} 0 \quad \textit{or equivalently} \quad d_2\bigl(\nu_n(v^0, v^1, \dots, v^k, v), \breve{\nu}^k\bigr) \xrightarrow{c} 0 \quad \textit{as } n \to \infty.$$

Before discussing Theorem 3.1 and its proof, we note that as an immediate consequence of (3.8), Corollary 3.2 below yields an exact expression for the asymptotic deviation of $\hat{v}^k = g_k(v^k)$ from v with respect to any pseudo-Lipschitz loss function of order 2. In particular, the asymptotic mean squared error and empirical correlation in (3.10) and (3.11) respectively depend only on λ and the state evolution parameters μ_{k+1}, σ_{k+1}.

Corollary 3.2. *In the setting of Theorem 3.1, fix $k \in \mathbb{N}$ and let $n \to \infty$. Then taking $G_k \sim N(0,1)$ to be independent of $V \sim \pi$, we have*

$$\frac{1}{n}\sum_{i=1}^n \psi(\hat{v}_i^k, v_i) \xrightarrow{c} \mathbb{E}\bigl\{\psi\bigl(g_k(\mu_k V + \sigma_k G_k), V\bigr)\bigr\} \tag{3.9}$$

for all $\psi \in \mathrm{PL}_2(2)$. Consequently,

$$\|\hat{v}^k - v\|_n^2 \xrightarrow{c} \mathbb{E}\bigl\{\bigl(g_k(\mu_k V + \sigma_k G_k) - V\bigr)^2\bigr\} = \sigma_{k+1}^2 - \frac{2\mu_{k+1}}{\lambda} + 1 \tag{3.10}$$

$$\textit{and} \quad \frac{|\langle \hat{v}^k, v\rangle_n|}{\|\hat{v}^k\|_n \|v\|_n} \xrightarrow{c} \frac{\bigl|\mathbb{E}\bigl(V g_k(\mu_k V + \sigma_k G_k)\bigr)\bigr|}{\sqrt{\mathbb{E}\bigl(g_k(\mu_k V + \sigma_k G_k)^2\bigr)}} = \frac{|\mu_{k+1}|}{\lambda \sigma_{k+1}}. \tag{3.11}$$

Remark 3.3. Observe that $\|v^k\|_n^2 \xrightarrow{c} \mathbb{E}\bigl((\mu_k V + \sigma_k G_k)^2\bigr) = \mu_k^2 + \sigma_k^2$ and $\|\hat{v}^{k-1}\|_n^2 = \|g_{k-1}(v^{k-1})\|_n \xrightarrow{c} \sigma_k^2$ for all $k \in \mathbb{N}$, so $\|v^k\|_n^2 - \|\hat{v}^{k-1}\|_n^2$ and $\|\hat{v}^{k-1}\|_n^2$ are strongly consistent estimators of μ_k^2 and σ_k^2 respectively.

Interpretation: Through the state evolution recursion (3.6), Corollary 3.2 establishes a precise correspondence between the asymptotic behaviour of the AMP iterates $\bigl(\hat{v}^k \equiv \hat{v}^k(n) : n \in \mathbb{N}\bigr)$ and a univariate deconvolution problem, where we estimate V by $g_k(\mu_k V + \sigma_k G_k)$ when given a single noisy observation $\mu_k V + \sigma_k G_k$. In this context, the quantity $\rho_k := (\mu_k/\sigma_k)^2$ can be interpreted as an *effective signal-to-noise ratio*, which arises naturally in (3.11) above. Returning to the spiked model (3.1), we can think of $\hat{v}^k \equiv \hat{v}^k(n) = g_k(v^k)$ as an estimate of $v \equiv v(n)$ based on an '*effective observation*' $v^k \equiv v^k(n)$ whose components have approximately the same empirical distribution as those of $\mu_k v + \sigma_k \xi$ when n is large, where $\xi \equiv \xi(n) \sim N_n(0, I_n)$ is independent of v.

Theorem 3.1 and Corollary 3.2 can be rigorously proved by means of an instructive application of the master theorems for the abstract symmetric AMP iteration (2.1) in Section 2.1. In the next few paragraphs (which can be skipped on a first reading), we will outline the key arguments in the setting of Corollary 3.2; a full proof of the more general Theorem 3.1 can be found in Section 6.8.

In summary, we begin by rewriting the AMP algorithm (3.3) in terms of the 'noise' components $\breve{u}^k \equiv \breve{u}^k(n) := v^k - \mu_k v$ of the effective observations v^k, and aim to show that the corresponding noise variables in the limiting univariate problem are indeed Gaussian (and independent of V), with mean 0 and variance σ_k^2 given by (3.6). To this end, it can be seen that the resulting recursion (3.12) below for $(\breve{u}^k : k \in \mathbb{N})$ is very similar to an iteration of the abstract form (2.1), whose exact asymptotics are given by Theorems 2.1 and 2.3. In addition to these main workhorse results, some additional technical arguments are needed to take care of a 'correction term' in (3.12) below with asymptotically vanishing influence.

The conclusion is that for each k, the joint empirical distribution $\nu_n(\breve{u}^k, v)$ of the entries of $\breve{u}^k(n) = v^k(n) - \mu_k v(n)$ and $v \equiv v(n)$ converges completely in $\widetilde{d}_2$ to the distribution of $(\sigma_k G_k, V)$ as $n \to \infty$. Equivalently, $\nu_n(v^k, v)$ converges completely in $\widetilde{d}_2$ to the distribution of $(\mu_k V + \sigma_k G_k, V)$ as $n \to \infty$, whence the conclusion (3.9) of Corollary 3.2 follows straightforwardly.

Proof sketch for Corollary 3.2. More precisely, under the spiked model (3.1), A is the sum of independent signal and noise matrices $\lambda v v^\top/n$ and W respectively, so (3.3) becomes $v^{k+1} \equiv v^{k+1}(n) = \lambda\langle \hat{v}^k, v\rangle_n v + W g_k(v^k) - b_k g_{k-1}(v^{k-1})$ for $k, n \in \mathbb{N}$. Rearranging this and defining

$$\delta_k \equiv \delta_k(n) := \lambda\langle \hat{v}^{k-1}, v\rangle_n - \mu_k$$

for all k and n, we see that $\breve{u}^k \equiv \breve{u}^k(n) = v^k(n) - \mu_k v(n)$ satisfies

$$\breve{u}^1 = W\hat{v}^0 + \delta_1 v, \qquad \breve{u}^{k+1} = W g_k(\breve{u}^k + \mu_k v) - b_k g_{k-1}(\breve{u}^{k-1} + \mu_{k-1} v) + \delta_{k+1} v \quad \text{for } k \in \mathbb{N}, \tag{3.12}$$

where $b_k \equiv b_k(n) = \langle g_k'(v^k)\rangle_n = \langle g_k'(\breve{u}^k + \mu_k v)\rangle_n$. Setting $u^1 := W\hat{v}^0$ and dropping the final $\delta_{k+1} v$ term from the right hand side of (3.12), we obtain a related recursion

$$u^{k+1} \equiv u^{k+1}(n) := W g_k(u^k + \mu_k v) - \tilde{b}_k g_{k-1}(u^{k-1} + \mu_{k-1} v) \quad \text{for } k \in \mathbb{N}, \tag{3.13}$$

where $\tilde{b}_k \equiv \tilde{b}_k(n) := \langle g_k'(u^k + \mu_k v)\rangle_n$. This is an instance of (2.1) with $f_k, f_k' \colon \mathbb{R}^2 \to \mathbb{R}$ given by $f_k(x, y) = g_k(x + \mu_k y)$ and $f_k'(x, y) = g_k'(x + \mu_k y)$ for $x, y \in \mathbb{R}$. Under (M0)–(M2), it is straightforward to verify that (A0)–(A5) are satisfied with

$$\tau_1^2 = \operatorname*{c-lim}_{n\to\infty} \|\hat{v}^0\|_n^2 = \sigma_1^2 \quad \text{and} \quad \tau_{k+1}^2 = \mathbb{E}\bigl(f_k(\sigma_k G_k, V)^2\bigr) = \mathbb{E}\bigl(g_k(\mu_k V + \sigma_k G_k)^2\bigr) = \sigma_{k+1}^2 \tag{3.14}$$

for all $k \in \mathbb{N}$ (by induction), in view of the state evolution recursion for $(\sigma_k : k \in \mathbb{N})$ in (3.5)–(3.6). It follows from Theorem 2.1 that for each k in (3.13), the joint empirical distribution $\nu_n(u^k, v)$ converges completely in d_2 to the distribution $N(0, \sigma_k^2) \otimes \pi$ of $(\sigma_k G_k, V)$ as $n \to \infty$.

It now remains to establish that the $\delta_{k+1} v$ term in (3.12) has asymptotically negligible effect, in the sense that the iterates in (3.12) remain close to those in (3.13) and hence have the same limiting distributions. Specifically, it can be shown by induction on $k \in \mathbb{N}$ that $\delta_k(n) \xrightarrow{c} 0$, that $\|\breve{u}^k - u^k\|_n \xrightarrow{c} 0$ and hence that $\widetilde{d}_2\bigl(\nu_n(\breve{u}^k, v), N(0, \sigma_k^2) \otimes \pi\bigr) \xrightarrow{c} 0$ as $n \to \infty$ for each fixed k. The arguments involved are fairly routine, and are spelled out in detail in Section 6.8. We mention here that the first part of the inductive step reveals the origins of the state evolution recursion for $(\mu_k : k \in \mathbb{N})$ in (3.5)–(3.6): it follows from the inductive hypothesis $\widetilde{d}_2\big(\nu_n(\breve{u}^k, v), N(0, \sigma_k^2) \otimes \pi\big) \xrightarrow{c} 0$ that

$$\lambda\langle \hat{v}^k, v\rangle_n = \frac{\lambda}{n}\sum_{i=1}^n v_i g_k(\breve{u}_i^k + \mu_k v_i) \xrightarrow{c} \lambda\mathbb{E}\big(V g_k(\mu_k V + \sigma_k G)\big) = \mu_{k+1}$$

as $n \to \infty$, so indeed $\delta_{k+1}(n) \xrightarrow{c} 0$ as $n \to \infty$. □

We conclude this subsection by noting that for a given sequence of (random) spikes $v \equiv v(n)$ satisfying (M1), the quality of the estimates $\hat{v}^k \equiv \hat{v}^k(n)$ clearly depends on the vectors $v^0 \equiv v^0(n)$ that are used to initiate the AMP iterations, as well as the sequence of Lipschitz functions $g_k \colon \mathbb{R} \to \mathbb{R}$. In the next two subsections, we will describe how these inputs to (3.3) can be suitably chosen to achieve good estimation performance, based on the information that we have about the distribution of V.

3.2 Spectral initialisation

In the context of the spiked model (3.1), it is helpful to think of AMP as a method by which we can potentially improve a 'pilot' estimator $\hat{v}^0 \equiv \hat{v}^0(n) = g_0(v^0)$ of $v \equiv v(n)$, in the sense that we may be able to increase the asymptotic empirical correlation in (3.11) (i.e. the effective signal-to-noise ratio) by repeatedly iterating (3.3). To this end, a minimum requirement is that we obtain effective signal-to-noise ratios ρ_{k+1} that are strictly positive, since the corresponding estimates $\hat{v}^k \equiv \hat{v}^k(n)$ ought to be at least partially aligned with v in the limit $n \to \infty$.

When $\mathbb{E}(V) \neq 0$, we will see in Section 3.3 that if the functions g_k are chosen appropriately, then it suffices to take $v^0 \equiv v^0(n) = c\mathbf{1}_n \equiv (c, \dots, c) \in \mathbb{R}^n$ for each n, where $c \in \mathbb{R}$ is fixed. However, this does not work when $\mathbb{E}(V) = 0$: in this case, $\mu_0 = \text{c-lim}_{n\to\infty}\langle c\mathbf{1}_n, v\rangle_n = 0$ in (M1), and for any choice of (g_k), the state evolution recursion (3.6) then yields $\mu_k = 0$ and $\rho_k = (\mu_k/\sigma_k)^2 = 0$ for all $k \in \mathbb{N}$ (since V and G_k are independent). For each k, it follows from (3.11) that $\langle \hat{v}^k, v\rangle_n \xrightarrow{c} 0$ as $n \to \infty$, so $\hat{v}^k \equiv \hat{v}^k(n)$ is asymptotically uninformative as an estimator of $v \equiv v(n)$.

Thus, when $\mathbb{E}(V) = 0$, we require $\mu_0 \neq 0$ and pilot estimators that have non-zero asymptotic empirical correlation with v. For $n \in \mathbb{N}$, consider initialising the AMP algorithm (3.3) with $v^0 = c\hat{\varphi}$ for some $c \neq 0$, where $\hat{\varphi} \equiv \hat{\varphi}(n)$ is a normalised principal eigenvector of $A \equiv A(n)$ with $\|\hat{\varphi}\|_n = 1$. This is almost surely well-defined up to its sign, and yields an initial estimate with the desired property precisely when $\lambda > 1$; indeed, recall from (3.2) that $|\langle \hat{\varphi}, v\rangle_n|/\|v\|_n \xrightarrow{c} \sqrt{1 - \lambda^{-2}} > 0$ for such λ. Using the orthogonal invariance of $W \sim \text{GOE}(n)$, Proposition 3.4 below extends this convergence result to show that $\big\{\big(\hat{\varphi}(n), v(n)\big) : n \in \mathbb{N}\big\}$ satisfies condition (M1) with $\mu_0 = \sqrt{1 - \lambda^{-2}}$, $\sigma_0 = 1/\lambda$ and $U \sim N(0, 1)$, provided that $\langle \hat{\varphi}, v\rangle_n \geq 0$ for all n; see Remark 3.6 below for further discussion of this final issue.

Proposition 3.4. *Suppose that $V \sim \pi$ satisfies $\mathbb{E}(V^2) = 1$ and $d_2\big(\nu_n(v), \pi\big) \xrightarrow{c} 0$ as $n \to \infty$, where $\nu_n(v) = n^{-1}\sum_{i=1}^n \delta_{v_i}$ denotes the empirical distribution of $v \equiv v(n)$ for $n \in \mathbb{N}$. If $\lambda > 1$ in (3.1), and each $\hat{\varphi} \equiv \hat{\varphi}(n)$ is a principal eigenvector of $A \equiv A(n)$ whose direction is chosen so that $\langle \hat{\varphi}, v\rangle_n \geq 0$ for all n, then*

$$\sup_{\psi \in \text{PL}_2(2,1)}\left|\frac{1}{n}\sum_{i=1}^n \psi(\hat{\varphi}_i, v_i) - \mathbb{E}\big\{\psi\big(\sqrt{1 - \lambda^{-2}}\,V + \lambda^{-1}G_0, V\big)\big\}\right| \xrightarrow{c} 0$$

as $n \to \infty$, where $G_0 \sim N(0, 1)$ is independent of V.

For proofs of more general results of this type for finite-rank perturbations of GOE matrices, see Montanari and Venkataramanan (2021, Lemma C.1 and Corollary C.3).

In the subsequent asymptotic analysis of the AMP algorithm (3.3) with spectral initialisation, an additional technical challenge stems from the fact that $\hat{\varphi} \equiv \hat{\varphi}(n)$ is not independent of the noise matrix

$W \equiv W(n)$ for any n. This means that condition (M0) does not hold in general, so the theory from Section 3.1 is not directly applicable in this setting. Nevertheless, Montanari and Venkataramanan (2021) established Theorem 3.5 below to recover the desired conclusion for this particular initialisation, with the same state evolution parameters μ_k, σ_k as defined in (3.6) but a slightly modified limiting covariance structure. For fixed $c \neq 0$, let $\mu_0 := c\sqrt{1-\lambda^{-2}}$ and $\bar{\Sigma}^{[0]} = \bar{\Sigma}_{0,0} \equiv \sigma_0^2 := c^2/\lambda^2$. For a general $k \in \mathbb{N}$, suppose inductively that we have already defined a non-negative definite $\bar{\Sigma}^{[k-1]} \in \mathbb{R}^{k \times k}$ with entries $\bar{\Sigma}^{[k-1]}_{ij} = \bar{\Sigma}_{i,j}$ for $0 \leq i, j \leq k-1$, and then let

$$\bar{\Sigma}_{k,\ell} = \bar{\Sigma}_{\ell,k} := \begin{cases} \lambda^{-1}\,\mathbb{E}\big((\mu_0 V + \sigma_0 G_0) \cdot g_{k-1}(\mu_{k-1}V + \sigma_{k-1}G_{k-1})\big) & \text{for } \ell = 0, \\ \mathbb{E}\big(g_{\ell-1}(\mu_{\ell-1}V + \sigma_{\ell-1}G_{\ell-1}) \cdot g_{k-1}(\mu_{k-1}V + \sigma_{k-1}G_{k-1})\big) & \text{for } \ell = 1, \dots, k, \end{cases} \tag{3.15}$$

where $(\sigma_0 G_0, \sigma_1 G_1, \dots, \sigma_{k-1}G_{k-1}) \sim N_k(0, \bar{\Sigma}^{[k-1]})$ is independent of $V \sim \pi$. Let $\bar{\Sigma}^{[k]}$ be the $(k+1) \times (k+1)$ matrix with entries $\bar{\Sigma}^{[k]}_{ij} := \bar{\Sigma}_{i,j}$ for $0 \leq i, j \leq k$. Similarly to (3.7), $\bar{\Sigma}^{[k]}$ is non-negative definite, and if we take $U \sim N(0,1)$ to be independent of $V \sim \pi$ in the state evolution recursion (3.6), then $\bar{\Sigma}_{k,k} = \sigma_k^2$ by induction. However, unlike in Section 3.1, observe from the first line of (3.15) that the limiting Gaussian variables $G_1, G_2, \dots$ need not be independent of $U \equiv G_0$. This reflects the dependence between W and the initialiser $v^0 = c\hat{\varphi}$ for each n. To ensure that subsequent AMP iterates in (3.3) have the correct asymptotics in this setting, we also set $v^{-1} = \lambda^{-1}c\,\hat{\varphi}$ instead of $v^{-1} = 0$.

Theorem 3.5. *Suppose that $\lambda > 1$ in the spiked model (3.1), and that the hypotheses of Proposition 3.4 are satisfied for a sequence of AMP algorithms (3.3) initialised with $v^0 \equiv v^0(n) = c\,\hat{\varphi}(n)$ and $\hat{v}^{-1} \equiv \hat{v}^{-1}(n) = \lambda^{-1}c\,\hat{\varphi}(n)$ for each $n \in \mathbb{N}$, where $\langle \hat{\varphi}, v\rangle_n \geq 0$ and $c \neq 0$ is fixed. Starting with $\mu_0 = c\sqrt{1-\lambda^{-2}}$, $\sigma_0 = c/\lambda$ and $U \equiv G_0 \sim N(0,1)$ in (3.5), define the state evolution parameters $\mu_k, \sigma_k, \bar{\Sigma}^{[k]}$ for $k \in \mathbb{N}$ according to (3.6) and (3.15). Then under (M2), the convergence results (3.8)–(3.11) remain valid.*

To circumvent the difficulty mentioned above, Theorem 3.5 can be proved by first applying the existing AMP machinery to a suitably modified version of the iteration (3.3) for which (M0) is satisfied, and then showing that this has the same asymptotics as the original procedure with spectral initialisation. In the spiked model (3.1) where the signal matrix has rank 1, one approach along these lines is to design a more tractable two-stage iteration, in which the input to (3.3) in the second phase is the output of a surrogate power method that approximates $v^0 = c\,\hat{\varphi}$. This 'artificial' first phase takes the form of an AMP iteration with specially chosen linear threshold functions (see (3.27) in Section 3.3) and a (non-spectral) initialiser that is independent of W. The success of this strategy relies on the fact that the spectral gap $\lambda_1(A) - \lambda_2(A)$ of $A \equiv A(n)$ has a strictly positive limit as $n \to \infty$ when $\lambda > 1$, as mentioned at the start of Section 3.1. For further details of applications of this proof technique in the GAMP setting of Section 4, see Mondelli et al. (2021) and Mondelli and Venkataramanan (2020).

We refer the reader to Montanari and Venkataramanan (2021, Appendix A) for a different proof of Theorem 3.5 that extends more readily to a wider class of AMP algorithms for general low-rank matrix estimation (see Section 3.5). This involves studying a variant of (3.3) in which $A \equiv A(n)$ is replaced with

$$\tilde{A} \equiv \tilde{A}(n) = \frac{\lambda_1(A)}{n}\hat{\varphi}\hat{\varphi}^\top + \hat{P}^\perp\Big(\frac{\lambda}{n}vv^\top + \tilde{W}\Big)\hat{P}^\perp$$

for each n, where $\lambda_1(A)$ is the maximal eigenvalue of A, the matrix $\hat{P} := I - \hat{\varphi}\hat{\varphi}^\top/n$ represents the projection onto the orthogonal complement of $\hat{\varphi}$, and (crucially) $\tilde{W} \sim \mathrm{GOE}(n)$ is independent of W and v. To relate the simplified iteration based on $\tilde{A}$ to the original AMP procedure, an important technical step is to show that the conditional distributions of A and $\tilde{A}$ given $\big(\hat{\varphi}, \lambda_1(A)\big)$ are close in total variation distance when n is large.

Remark 3.6. In an estimation context where each $v \equiv v(n)$ is unknown, it is sometimes not possible to consistently determine the sign of the leading eigenvector of $A \equiv A(n)$ that should be used as a spectral initialiser, to ensure that it has non-negative asymptotic empirical correlation with v. For example, this is the case if the limiting prior distribution π is symmetric, i.e. $V \overset{d}{=} -V$. On the event of probability 1 where A has a unique maximal eigenvalue, suppose that one of the two possible directions for the corresponding eigenvector is chosen uniformly at random when carrying out spectral initialisation. In other words, let $v^0 \equiv v^0(n) = \epsilon\hat{\varphi}$ for each n, where $\langle \hat{\varphi}, v\rangle_n \geq 0$ and $\epsilon \equiv \epsilon(n)$ is a Rademacher random variable that is independent of everything else. With this choice of v^0, there are two different state evolution trajectories that can arise: for $\epsilon' \in \{-1, 1\}$, let $\mu_0(\epsilon') := \epsilon'\sqrt{1-\lambda^{-2}}$ and σ_0, U be as above, and define $\mu_k(\epsilon'), \sigma_k(\epsilon'), \bar{\Sigma}^{[k]}(\epsilon')$ for $k \in \mathbb{N}$ as per (3.6)–(3.7). For the resulting iterates $v^k \equiv v^k(n)$, Theorem 3.5 implies that as $n \to \infty$, we have

$$\sup_{\psi \in \mathrm{PL}_2(2,1)} \left| \frac{1}{n}\sum_{i=1}^n \psi(v_i^k, v_i) - \mathbb{E}\bigl\{\psi\bigl(\mu_k(\epsilon)V + \sigma_k(\epsilon)G_k, V\bigr) \bigm| \epsilon\bigr\}\right| \xrightarrow{c} 0$$

for each $k \in \mathbb{N}_0$, as well as appropriate analogues of (3.8)–(3.11). Since $\mathbb{E}\bigl\{\psi\bigl(\mu_k(\epsilon)V+\sigma_k(\epsilon)G_k, V\bigr) \bigm| \epsilon\bigr\}$ is random for each ψ, the empirical distribution of the components of $v^k(n)$ may not converge (completely in d_2) to a deterministic limit as $n \to \infty$, unlike in earlier results. Instead, we see that for large n, the behaviour of the AMP iterates is characterised by a state evolution recursion with a random initial condition $\mu_0(\epsilon)$ that depends on $v^0 \equiv v^0(n)$ through the (unknown) sign $\epsilon \equiv \epsilon(n)$.

3.3 Choosing the functions g_k

Recall that our goal is to specialise the general AMP algorithm (3.3) to produce estimates $\hat{v}^k = g_k(v^k)$ of v that exploit full or partial knowledge of the limiting prior distribution π from (M1). Corollary 3.2 suggests that we should aim to choose a sequence of Lipschitz ‘denoising’ functions $g_k\colon \mathbb{R} \to \mathbb{R}$ for which each $g_k(\mu_k V + \sigma_k G_k)$ performs well as an estimator of $V \sim \pi$ in the limiting univariate problem, where $G_k \sim N(0,1)$ is independent of V for $k \in \mathbb{N}$. More precisely, it would be desirable to ensure that the effective signal-to-noise ratio $\rho_k = (\mu_k/\sigma_k)^2$ is large for each k, since (3.11) tells us that the asymptotic empirical correlation between $\hat{v}^k$ and v is given by $\sqrt{\rho_{k+1}}/\lambda$. In fact, the implication of Lemma 3.7 below is that achieving a high effective signal-to-noise ratio ought to be our first priority, even when the ultimate objective is for $\hat{v}^k = g_k(v^k)$ to have low asymptotic estimation error $\mathbb{E}\bigl\{\psi\bigl(g_k(\mu_k V+\sigma_k G_k), V\bigr)\bigr\}$ with respect to some specific loss function $\psi \in \mathrm{PL}_2(2)$.

Lemma 3.7. *Let $G \sim N(0,1)$ be independent of $V \sim \pi$. Then for any Borel measurable loss function $\psi\colon \mathbb{R}^2 \to [0,\infty)$,*

$$\rho \mapsto \inf_g \mathbb{E}\bigl\{\psi\bigl(g(\sqrt{\rho}V + G), V\bigr)\bigr\} =: R_{\pi,\psi}(\rho)$$

is non-increasing on $[0,\infty)$, where the infimum is over all Borel measurable functions $g\colon \mathbb{R} \to \mathbb{R}$. This infimum is attained for all $\rho \in [0,\infty)$ if for example $\psi(x,y) = \Psi(x-y)$ for some convex function Ψ with $\Psi(u) \to \infty$ as $|u| \to \infty$.

The intuition behind this result is straightforward: to minimise $\mathbb{E}\bigl\{\psi\bigl(g(\sqrt{\rho}V + G), V\bigr)\bigr\}$ jointly over ρ (belonging to a given range) and all measurable g, we should always begin by taking the largest possible ρ (i.e. the least noisy $\sqrt{\rho}V + G$) before subsequently optimising over g. A formal proof of Lemma 3.7 is deferred to Section 6.8. The arguments therein show also that the first assertion of the lemma remains valid if the infimum is instead taken only over Lipschitz functions (which are more relevant to the setting of AMP).

Note that for (known) $\mu, \sigma \in \mathbb{R}$ with $(\mu/\sigma)^2 = \rho$, the quantity $R_{\pi,\psi}(\rho)$ is the π-Bayes risk with respect to ψ in a Bayesian mean estimation problem where we place a prior π on V and observe $Y = \mu V + \sigma G$ (as in the paragraph above), i.e. $Y \,|\, V \sim N(\mu V, \sigma^2)$. If there exists a Borel measurable $g^*\colon \mathbb{R} \to \mathbb{R}$ that attains the infimum in the definition of $R_{\pi,\psi}(\rho)$, then $g^*(Y)$ is a π-Bayes estimator of V (with respect to ψ) based on Y.

Bayes-AMP: Suppose first that for some $k \in \mathbb{N}_0$, we are given the distribution π of V and the state evolution parameters μ_k, σ_k (which depend on μ_0, σ_0 in (M1) as well as the functions $g_0, g_1, \ldots, g_{k-1}$).

For convenience, when $k = 0$, we write G_0 for the random variable U from (M1), and assume that its distribution is also known. Let $g_k^*\colon \mathbb{R} \to \mathbb{R}$ be any measurable function with

$$g_k^*(\mu_k V + \sigma_k G_k) = \mathbb{E}(V \,|\, \mu_k V + \sigma_k G_k), \tag{3.16}$$

which in principle can be computed based on $Y_k := \mu_k V + \sigma_k G_k$. In particular, for $k \in \mathbb{N}$, we have $G_k \sim N(0,1)$, in which case if $\sigma_k > 0$, then Y_k has a smooth (real analytic), strictly positive Lebesgue density p_k on $\mathbb{R}$. Specifically, $p_k(y) := \int_{\mathbb{R}} \phi_{\sigma_k}(y - \mu_k x)\, d\pi(x)$ for $y \in \mathbb{R}$, where ϕ_{σ_k} is the density of a $N(0, \sigma_k^2)$ random variable. Then by *Tweedie's formula* (Robbins, 1956; Efron, 2011), we can take

$$g_k^*(y) = \frac{y + \sigma_k^2\,(\log p_k)'(y)\mathbb{1}_{\{\sigma_k \neq 0\}}}{\mu_k}\mathbb{1}_{\{\mu_k \neq 0\}} = \frac{y + \sigma_k^2\,(p_k'/p_k)(y)\mathbb{1}_{\{\sigma_k \neq 0\}}}{\mu_k}\mathbb{1}_{\{\mu_k \neq 0\}} \quad \text{for } y \in \mathbb{R}. \tag{3.17}$$

For example, if π is the uniform distribution on $\{-1, 1\}$ and $\sigma_k \neq 0$, then $g_k^*(y) = \tanh(\mu_k y/\sigma_k^2)$ for $y \in \mathbb{R}$.

In the AMP literature, g_k^* is referred to as the 'Bayes optimal' choice of threshold function in (3.3), since the posterior mean $g_k^*(Y_k) = \mathbb{E}(V \,|\, Y_k)$ is the Bayes estimator of V based on Y_k with respect to quadratic loss (often known as the *minimum mean squared error* (MMSE) estimator). Indeed, by the characterisation of $\mathbb{E}(V \,|\, Y_k)$ as an orthogonal projection,

$$\mathbb{E}\bigl\{\bigl(V - g(Y_k)\bigr)^2\bigr\} = \mathbb{E}\bigl\{\bigl(V - g_k^*(Y_k)\bigr)^2\bigr\} + \mathbb{E}\{(g_k^* - g)^2(Y_k)\} \geq \mathbb{E}\bigl\{\bigl(V - g_k^*(Y_k)\bigr)^2\bigr\} \tag{3.18}$$

for all measurable $g\colon \mathbb{R} \to \mathbb{R}$. In addition,

$$\frac{\mathbb{E}\bigl(V g(Y_k)\bigr)^2}{\mathbb{E}\bigl(g(Y_k)^2\bigr)} = \frac{\mathbb{E}\bigl(g_k^*(Y_k)\, g(Y_k)\bigr)^2}{\mathbb{E}\bigl(g(Y_k)^2\bigr)} \leq \mathbb{E}\bigl(g_k^*(Y_k)^2\bigr) \tag{3.19}$$

by the Cauchy–Schwarz inequality, with equality if g is a (non-zero) scalar multiple of g_k^*. Thus, for given μ_k, σ_k, the function g_k^* simultaneously minimises the asymptotic mean squared error in (3.10) and maximises the asymptotic empirical correlation (i.e. the effective signal-to-noise ratio ρ_{k+1}^*) in (3.11) over all measurable $g_k\colon \mathbb{R} \to \mathbb{R}$.

The following result is a slight extension of Montanari and Venkataramanan (2021, Remark 2.3) (with a different, simpler proof given in Section 6.8) that provides sufficient conditions on π under which g_k^* is Lipschitz and satisfies (M2).

Lemma 3.8. *Suppose either that V has a log-concave density, or that there exist independent random variables U_0, V_0 such that U_0 is Gaussian, V_0 is compactly supported and $V \overset{d}{=} U_0 + V_0$. Then for $\mu_k, \sigma_k \neq 0$, the function g_k^* in* (3.17) *is smooth and Lipschitz on $\mathbb{R}$.*

Assuming now that we have complete knowledge of the distributions of U, V as well as $\lambda > 0$ in (3.1) and μ_0, σ_0 from (M1), we can construct a *'Bayes-AMP' algorithm* of the form (3.3) by recursively defining $(g_k^* : k \in \mathbb{N})$ and state evolution sequences $(\mu_k^*, \sigma_k^* : k \in \mathbb{N})$ in accordance with (3.16, 3.17) and (3.6) respectively. We will write $v^{k,\mathrm{B}} \equiv v^{k,\mathrm{B}}(n)$ for the resulting Bayes-AMP iterates (i.e. effective observations) and $\hat{v}^{k,\mathrm{B}} \equiv \hat{v}^{k,\mathrm{B}}(n) := g_k^*(v^{k,\mathrm{B}})$ for the Bayes-AMP estimates of $v \equiv v(n)$.

For each $k \in \mathbb{N}$, we have $\mu_{k+1}^* = \lambda\mathbb{E}\bigl(V g_k^*(Y_k)\bigr) = \lambda\mathbb{E}\bigl(g_k^*(Y_k)^2\bigr) = \lambda(\sigma_{k+1}^*)^2$ by (3.19), and since $\mathbb{E}(V^2) = 1$ by (M1), the effective signal-to-noise ratios in Bayes-AMP satisfy

$$\rho_{k+1}^* := (\mu_{k+1}^*/\sigma_{k+1}^*)^2 = \lambda^2(\sigma_{k+1}^*)^2 = \lambda^2\,\mathbb{E}\bigl(g_k^*(Y_k)^2\bigr) = \lambda^2\bigl(1 - \mathbb{E}\bigl\{\bigl(V - g_k^*(Y_k)\bigr)^2\bigr\}\bigr). \tag{3.20}$$

Thus, the state evolution recursion (3.6) for Bayes-AMP can be compactly written as

$$\rho_0^* := (\mu_0/\sigma_0)^2, \quad \rho_{k+1}^* := \lambda^2\bigl(1 - \mathrm{mmse}_k(\rho_k^*)\bigr) \quad \text{for } k \in \mathbb{N}_0, \tag{3.21}$$

where for $\rho \in [0, \infty)$ we denote by

$$\mathrm{mmse}_k(\rho) := \mathbb{E}\bigl\{\bigl(V - \mathbb{E}(V \,|\, \sqrt{\rho}V + G_k)\bigr)^2\bigr\}$$

the *minimum mean squared error* (i.e. the Bayes risk with respect to squared error loss $\psi_2 \colon (x,y) \mapsto (x-y)^2$) for the problem of reconstructing V based on the corrupted observation $\sqrt{\rho}V + G_k$. For $k \in \mathbb{N}$, we have $G_k \sim N(0,1)$, in which case we simply write $\mathrm{mmse}(\rho)$ for $\mathrm{mmse}_k(\rho) = R_{\pi,\psi_2}(\rho)$. For concreteness, we set $\mathrm{mmse}(\infty) = 0$, which is consistent with the fact that $\mathrm{mmse}(\rho) \to 0$ as $\rho \to \infty$.

At each iteration $k \in \mathbb{N}$, it turns out that ρ_{k+1}^* is the highest effective signal-to-noise ratio that can be achieved with any choice of functions (g_k) in the generic AMP procedure (3.3).

Corollary 3.9. *Consider any sequence of AMP iterations $\bigl(v^k \equiv v^k(n) : k, n \in \mathbb{N}\bigr)$ of the form (3.3) for which the hypotheses of Theorem 3.1 or 3.5 are satisfied with $V \sim \pi$ and suitable μ_0, σ_0. Let $(\mu_k, \sigma_k : k \in \mathbb{N}_0)$ and $\bigl(\rho_k = (\mu_k/\sigma_k)^2 : k \in \mathbb{N}_0\bigr)$ be the associated sequences of state evolution parameters and effective signal-to-noise ratios respectively. Define $(\rho_k^* : k \in \mathbb{N}_0)$ as in (3.20). Then for each $k \in \mathbb{N}_0$ and any $\psi \in \mathrm{PL}_2(2)$, the estimates $\hat{v}^k \equiv \hat{v}^k(n) = g_k(v^k)$ satisfy*

$$\frac{|\langle \hat{v}^k, v\rangle_n|}{\|\hat{v}^k\|_n \|v\|_n} \xrightarrow{c} \frac{\sqrt{\rho_{k+1}}}{\lambda} \leq \frac{\sqrt{\rho_{k+1}^*}}{\lambda} \tag{3.22}$$

$$\text{and} \qquad \frac{1}{n}\sum_{i=1}^n \psi(\hat{v}_i^k, v_i) \xrightarrow{c} \mathbb{E}\bigl\{\psi\bigl(g_k(\mu_k V + \sigma_k G_k), V\bigr)\bigr\} \geq R_{\pi,\psi}(\rho_k^*) \quad \text{as } n \to \infty. \tag{3.23}$$

This follows from (3.11) and (3.19) above, as well as Lemma 3.7, which implies in particular that $\rho \mapsto \mathrm{mmse}(\rho)$ is decreasing on $[0,\infty)$. See Section 6.8 for a full justification of Corollary 3.9.

Under the conditions of Lemma 3.8 above, the Bayes optimal functions g_k^* are Lipschitz and satisfy (M2). We can then apply the general results in Sections 3.1 and 3.2 to obtain the exact asymptotics for Bayes-AMP, for which it follows that (3.22) holds with equality. In other words, at every iteration, the Bayes-AMP estimate $\hat{v}^{k,\mathrm{B}} = g_k^*(v^{k,\mathrm{B}})$ achieves the optimal asymptotic empirical correlation among all AMP algorithms that are covered by the theory above. Moreover, with the initialisations in (i) and (ii) below, Theorem 3.10 shows that Bayes-AMP achieves the objective set out at the start of Section 3.2, namely that $\hat{v}^{k+1,\mathrm{B}}$ is a strict improvement on $\hat{v}^{k,\mathrm{B}}$ in terms of its asymptotic squared error and empirical correlation (i.e. the effective signal-to-noise ratio ρ_{k+1}^*) for each k. This means that for large k and n, the performance of $\hat{v}^{k,\mathrm{B}} \equiv \hat{v}^{k,\mathrm{B}}(n)$ is approximately characterised by a fixed point of the recursion in (3.21) to which (ρ_k^*) converges monotonically; see Figure 2.

Theorem 3.10. *Let $\bigl(v^{k,\mathrm{B}} \equiv v^{k,\mathrm{B}}(n) : k, n \in \mathbb{N}\bigr)$ be a sequence of Bayes-AMP iterations that satisfies either (i) or (ii) below.*

(i) (Non-spectral initialisation) $v^0 \equiv v^0(n) = c\mathbf{1}_n$ *and* $v^{-1} \equiv v^{-1}(n) = 0$ *for each n, where $c \in \mathbb{R}$ is fixed. Suppose that the hypotheses of Theorem 3.1 are satisfied with $\mathbb{E}(V) \neq 0$, in which case $\mu_0 = 0$, $\sigma_0 = c$ and $\rho_0^* = 0$.*

(ii) (Spectral initialisation) $v^0 \equiv v^0(n) = c\,\hat{\varphi}(n)$ *and* $v^{-1} \equiv v^{-1}(n) = \lambda^{-1}c\,\hat{\varphi}(n)$ *for each n, where $c \neq 0$ is fixed and $\langle \hat{\varphi}, v\rangle_n \geq 0$. Suppose that the hypotheses of Theorem 3.5 are satisfied with $\lambda > 1$, in which case $\mu_0 = c\sqrt{1-\lambda^{-2}}$, $\sigma_0 = c/\lambda$ and $\rho_0^* = \lambda^2 - 1$.*

Suppose that $V \sim \pi$ satisfies one of the conditions of Lemma 3.8. Then we have the following:

(a) The sequence $(\rho_k^* : k \in \mathbb{N}_0)$ of effective signal-to-noise ratios defined through (3.20) is strictly increasing, and converges to the smallest *strictly positive fixed point of $\rho = \lambda^2\bigl(1 - \mathrm{mmse}(\rho)\bigr)$, which we denote by $\rho_{\mathrm{AMP}}^* \equiv \rho_{\mathrm{AMP}}^*(\lambda) \in (0, \lambda^2]$.*

(b) For $k \in \mathbb{N}$ and a (convex, non-negative) loss function $\psi \in \mathrm{PL}_2(2)$, suppose that $g_{k,\psi}^* \colon \mathbb{R} \to \mathbb{R}$ is Lipschitz and attains the infimum in the definition of $R_{\pi,\psi}(\rho_k^*)$. Then the estimates $\hat{v}^{k,\psi} \equiv \hat{v}^{k,\psi}(n) := g_{k,\psi}^*(v^{k,\mathrm{B}})$ satisfy (3.23) with equality, i.e. $n^{-1}\sum_{i=1}^n \psi(\hat{v}_i^{k,\psi}, v_i) \xrightarrow{c} R_{\pi,\psi}(\rho_k^*)$ as $n \to \infty$, and $R_{\pi,\psi}(\rho_k^*) \geq R_{\pi,\psi}(\rho_{k+1}^*)$.

(c) The Bayes-AMP estimates $\hat{v}^{k,\mathrm{B}} = g_k^*(v^{k,\mathrm{B}})$ satisfy

$$\operatorname*{c-lim}_{n\to\infty} \|\hat{v}^{k,\mathrm{B}} - v\|_n^2 = 1 - \frac{\rho_{k+1}^*}{\lambda^2} \searrow 1 - \frac{\rho_{\mathrm{AMP}}^*(\lambda)}{\lambda^2} \tag{3.24}$$

$$\text{and} \qquad \operatorname*{c-lim}_{n\to\infty} \frac{\langle \hat{v}^{k,\mathrm{B}}, v\rangle_n}{\|\hat{v}^{k,\mathrm{B}}\|_n \|v\|_n} = \frac{\sqrt{\rho_{k+1}^*}}{\lambda} \nearrow \frac{\sqrt{\rho_{\mathrm{AMP}}^*(\lambda)}}{\lambda} \qquad \text{as } k \to \infty. \tag{3.25}$$

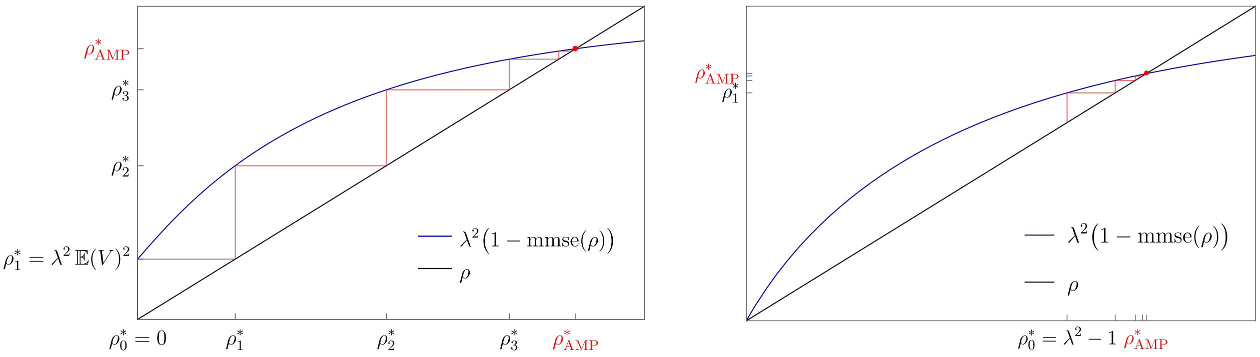

Figure 2: 'Cobweb diagrams' illustrating the conclusion of Theorem 3.10(a) that $\rho_k^* \nearrow \rho_{\mathrm{AMP}}^*$ as $k \to \infty$, under (i) and (ii) respectively with $\lambda = 1.7$; note that $1 - \mathrm{mmse}(0) = \mathbb{E}(V)^2$:

Left, non-spectral initialisation: $V \sim \pi = \frac{3}{4}\delta_0 + \frac{1}{4}\delta_2$, with $\mathbb{E}(V) = 1/2 \neq 0$ and $\mathbb{E}(V^2) = 1$: convergence to ρ_{AMP}^* occurs when $\rho_0^* = 0$.

Right, spectral initialisation: $V \sim \pi = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$, with $\mathbb{E}(V) = 0$ and $\mathbb{E}(V^2) = 1$: convergence to ρ_{AMP}^* occurs only if $\rho_0^* > 0$.

The proof of Theorem 3.10 is given in Section 6.8. To understand the implications of (b) above, suppose that we wish to use AMP to obtain estimates $\tilde{v} \equiv \tilde{v}(n)$ of $v \equiv v(n)$ with small (asymptotic) ℓ_1 estimation error $n^{-1}\sum_{i=1}^n |\tilde{v}_i - v_i|$. In view of (3.23) and Theorem 3.10(b), with ψ taken to be absolute error loss $\psi_1\colon (x,y) \mapsto |x-y|$, we should first run Bayes-AMP to obtain the highest possible effective signal-to-noise ratio ρ_k^* at every iteration. Then for each $k \in \mathbb{N}$, we should consider $g_{k,\psi_1}^*\colon \mathbb{R} \to \mathbb{R}$ for which $g_{k,\psi_1}^*(y)$ is a median of the conditional (i.e. posterior) distribution of V given $\mu_k V + \sigma_k G = y$ for (Lebesgue almost) every $y \in \mathbb{R}$. If we can find a Lipschitz g_{k,ψ_1}^* with this property, then $\hat{v}^{k,\psi_1} := g_{k,\psi_1}^*(v^{k,\mathrm{B}})$ attains the lowest possible limiting mean absolute error $R_{\pi,\psi_1}(\rho_k^*) = \inf_g \mathbb{E}\bigl\{\bigl|V - g\bigl(\sqrt{\rho_k^*}\,V + G\bigr)\bigr|\bigr\}$, among all estimators obtained from the k^{th} iteration of some AMP algorithm of the form (3.3). In cases where there is no suitable Lipschitz g_{k,ψ_1}^*, for example when V has a discrete distribution, one possible modification of the approach above would be to replace g_{k,ψ_1}^* with a Lipschitz approximation when constructing the estimator, in the hope that the resulting asymptotic ℓ_1 error is close to $R_{\pi,\psi_1}(\rho_k^*)$.

As for Theorem 3.10(c), one can compare the asymptotic mean squared error (3.24) and empirical correlation (3.25) achieved by Bayes-AMP with the corresponding Bayes optimal quantities (i.e. the best possible limiting values that can be attained by *any* estimator). In a spiked model (3.1) where the entries of $v \equiv v(n)$ are i.i.d. with distribution π, closed-form asymptotic expressions for the Bayes estimator $\mathbb{E}(v \,|\, A)$ were rigorously established by Barbier et al. (2016) and Lelarge and Miolane (2019). It turns out that the Bayes optimal performance is characterised by a fixed point ρ_{B}^* of $\rho = \lambda^2\bigl(1 - \mathrm{mmse}(\rho)\bigr)$ that maximises a specific free-energy functional; see Montanari and Venkataramanan (2021, Section 2.4) for further details. Thus, we can precisely characterise the performance gap between Bayes-AMP and Bayes optimal estimation for symmetric rank-one matrix estimation. In particular, when the equation $\rho = \lambda^2\bigl(1 - \mathrm{mmse}(\rho)\bigr)$ has a unique positive solution (as is the case for the $U\{-1,1\}$ prior in Figure 1), Bayes-AMP achieves the Bayes optimal performance. Furthermore, in cases where $\rho_{\mathrm{AMP}}^* \neq \rho_{\mathrm{B}}^*$ (i.e. AMP is not Bayes optimal), there is currently no known polynomial-time algorithm that is superior to Bayes-AMP in terms of the limiting effective signal-to-noise ratio in (3.25).

We remark that the optimality of Bayes-AMP among polynomial-time algorithms is conjectured only for certain classes of statistical problems such as low-rank matrix estimation and generalised linear models. For tensor PCA, Hopkins et al. (2015) proposed a polynomial-time algorithm based on the sum-of-squares hierarchy that has strictly better estimation performance than AMP (Montanari and Richard, 2014). Subsequently, Wein et al. (2019) developed generalised spectral methods based on statistical physics that match the performance of the sum-of-squares algorithm for tensor PCA.

Remark 3.11. Suppose that the limiting prior distribution π is symmetric, i.e. $V \overset{d}{=} -V$, in which case v and $-v$ are asymptotically indistinguishable. Then $\mathbb{E}(V) = 0$, and as mentioned in Remark 3.6, it is not possible to consistently choose the sign of the spectral initialiser in a data-driven way, so as to ensure that $\langle \hat{\varphi}, v \rangle_n \geq 0$ for each n. Nevertheless, the two possible state evolution trajectories for Bayes-AMP (with spectral initialisation) are easily seen to be identical up to the sign of each μ_k, so the limits in (3.24) and (3.25) remain valid for

$$\min_{\epsilon \in \{-1,1\}} \|\hat{v}^{k,\mathrm{B}} - \epsilon v\|_n^2 \qquad \text{and} \qquad \min_{\epsilon \in \{-1,1\}} \frac{\langle \hat{v}^{k,\mathrm{B}}, \epsilon v \rangle_n}{\|\hat{v}^{k,\mathrm{B}}\|_n \|v\|_n} \qquad \text{respectively.}$$

Remark 3.12. If the limiting prior distribution π is known but some or all of λ, μ_0, σ_0 are not, then starting with $\hat{v}^0 \equiv \hat{v}^0(n)$ for some n, we can construct an '*empirical Bayes-AMP algorithm*' based on estimates of μ_k, σ_k for each k. Specifically, recalling Remark 3.3 and proceeding inductively, we can use (3.17) to define $\hat{g}_k^*$ based on

$$\hat{\mu}_k := \bigl(\|v^k\|_n^2 - \|\hat{v}^{k-1}\|_n^2\bigr)^{1/2} \quad \text{and} \quad \hat{\sigma}_k := \|\hat{v}^{k-1}\|_n,$$

and then obtain $\hat{v}^k = \hat{g}_k^*(v^k)$ and v^{k+1} via (3.3) for each k. Alternatively, since $\mu_k^* = \lambda(\sigma_k^*)^2 \geq 0$ in Bayes-AMP, we could instead take $\hat{\mu}_k = \hat{\lambda}\hat{\sigma}_k^2 = \hat{\lambda}\|\hat{v}^{k-1}\|_n^2$, where

$$\hat{\lambda} := \frac{\lambda_1(A) + \sqrt{\lambda_1(A)^2 - 4}}{2}$$

is a strongly consistent estimator of λ by (3.2). Yet another approach is to first define $(\hat{\rho}_k)$ recursively by $\hat{\rho}_1 := \hat{\mu}_1^2/\hat{\sigma}_1^2$ and $\hat{\rho}_{k+1} := \hat{\lambda}^2\bigl(1 - \mathrm{mmse}(\hat{\rho}_k)\bigr)$ for each k. In view of (3.20), we can then estimate μ_k, σ_k^2 by $\hat{\rho}_{k+1}/\hat{\lambda}$ and $\hat{\rho}_{k+1}/\hat{\lambda}^2$ respectively, and use these to define $\hat{g}_k^*$ and hence $\hat{v}^k, v^{k+1}$ for each k as above. The theoretical guarantees in Theorems 3.1 and 3.5 extend fairly straightforwardly to empirical Bayes-AMP; see Montanari and Venkataramanan (2021, Lemma G.1).

Sparse signal recovery: To give another example where the AMP procedure (3.3) can be specialised appropriately, suppose that the exact distribution π of V is not known, but that for some fixed $s \in (0,1)$ and every $n \in \mathbb{N}$, the spike $v \equiv v(n)$ is known to have at most sn non-zero entries. This implies that π satisfies $\pi(\{0\}) \geq 1 - s$. In line with the classical theory on denoising sparse vectors (Donoho and Johnstone, 1994, 1998; Montanari, 2012, Section 9.3), we can take $(g_k)_{k \in \mathbb{N}_0}$ to be a sequence of soft-thresholding functions

$$g_k(y) = \mathrm{ST}_{t_k}(y) := \mathrm{sgn}(y)(|y| - t_k)_+,$$

so that the AMP algorithm (3.3) becomes

$$\hat{v}^k = \mathrm{ST}_{t_k}(v^k), \qquad b_k = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\hat{v}_i^k \neq 0\}} \equiv \frac{\|\hat{v}^k\|_0}{n}, \qquad v^{k+1} = A\hat{v}^k - b_k\hat{v}^{k-1} \qquad \text{for } k \in \mathbb{N}_0. \tag{3.26}$$

When each of the thresholds $t_k \in (0, \infty)$ is suitably chosen in terms of λ and the state evolution parameter σ_k (or consistent estimators thereof), Montanari and Venkataramanan (2021) establish lower bounds on the effective signal-to-noise ratios $\rho_k = (\mu_k/\sigma_k)^2$ that hold uniformly over the class of distributions π with $\pi(\{0\}) \geq 1 - s$. In conjunction with Corollary 3.2, this analysis leads to a theoretical guarantee on the performance of (3.26) for any sequence of ns-sparse spikes $v \equiv v(n)$ satisfying (M1); see Proposition 2.1 in the aforementioned paper.

We also mention that Barbier et al. (2020) recently established statistical and computational limits for sparse signal recovery in an asymptotic regime where the expected number of non-zero entries of

$v \equiv v(n)$ is a sublinear function of n. Specifically, for each n, the entries of v are drawn independently from a prior π_n with $\pi_n(\{0\}) \geq 1 - s_n$ and $s_n \to 0$ as $n \to \infty$. In this setting, the analysis makes use of finite-sample versions of the AMP master theorems (see Remark 6.3).

In summary, the state evolution characterisation of the AMP algorithm (3.3) allows us to choose the functions g_k in a principled way, depending on the prior information available about the signal v. A poor choice of (g_k) will lead to low effective signal-to-noise ratios $\rho_k = (\mu_k/\sigma_k)^2$, but the asymptotic convergence results (3.8)–(3.11) will continue to hold provided that the hypotheses of Theorem 3.1 or 3.5 are satisfied.

A key strength of the AMP framework is that it gives us the flexibility to choose non-linear functions g_k, such as the soft-thresholding functions above. Note that the MMSE denoising functions g_k^* in (3.16) are non-linear except in special cases (such as when V is Gaussian). Nevertheless, iterations with linear g_k can sometimes be useful as a theoretical device for obtaining distributional information about spectral estimators, as the following example shows; see also Mondelli and Venkataramanan (2020) and Mondelli et al. (2021).

Connection with the power method: Suppose that we initialise (3.3) with $\hat{v}^0 \equiv \hat{v}^0(n) := \mu_0 v + \xi$, where $\mu_0 \neq 0$ and $\xi \equiv \xi(n) \sim N_n(0, I_n)$, and define

$$\beta_k := \sqrt{1 + \mu_k^2}, \qquad g_k(x) := \frac{x}{\beta_k} \quad \text{for } x \in \mathbb{R}, \qquad \mu_{k+1} := \frac{\lambda}{\sqrt{1 + \mu_k^{-2}}} \quad \text{for } k \in \mathbb{N}_0. \tag{3.27}$$

These functions g_k are constructed in a such a way that the corresponding state evolution formula (3.6) yields $\sigma_k^2 = 1$ for every k, and parameters μ_k that coincide exactly with those defined in (3.27). Observe now that the AMP iteration (3.3) corresponding to (3.27) yields $\bigl(\hat{v}^k \equiv \hat{v}^k(n) : k \in \mathbb{N}\bigr)$ satisfying $\beta_0 \hat{v}^1 = v^1 = A\hat{v}^0$ and

$$\left(\beta_k + \frac{1}{\beta_{k-1}}\right)\hat{v}^{k+1} - \frac{1}{\beta_{k-1}}(\hat{v}^{k+1} - \hat{v}^{k-1}) = A\hat{v}^k \quad \text{for } k \in \mathbb{N}. \tag{3.28}$$

The key steps in the theoretical analysis of (3.28) can be summarised as follows:

(I) When $\lambda > 1$, some elementary analysis (e.g. based on the contraction mapping theorem) shows that $\sqrt{\lambda^2 - 1}$ is a stable fixed point of the deterministic recursion for (μ_k) in (3.27), and hence that $\beta_k \to \lambda$ as $k \to \infty$.

(II) Using Theorem 3.1 and the covariance matrix defined in (3.7), we can obtain the d_2 limit of the joint empirical distribution of the components of $\hat{v}^{k+1} \equiv \hat{v}^{k+1}(n)$ and $\hat{v}^{k-1} \equiv \hat{v}^{k-1}(n)$ as $n \to \infty$; in particular, $\|\hat{v}^{k+1}\|_n \xrightarrow{c} 1$. It then follows from (I) and routine arguments that $\lim_{k\to\infty} \text{c-lim}_{n\to\infty} \|\hat{v}^{k+1} - \hat{v}^{k-1}\|_n = 0$. In other words, $\|\hat{v}^{k+1} - \hat{v}^{k-1}\|_n$ converges completely to some deterministic limit ℓ_k as $n \to \infty$ for each fixed k, and $\ell_k \to 0$ as $k \to \infty$.

(III) Thus, writing (3.28) in the form $A\hat{v}^k = (\lambda + \lambda^{-1})\hat{v}^{k+1} + \vartheta_k$ for $k, n \in \mathbb{N}$, where

$$\vartheta_k \equiv \vartheta_k(n) := \left\{\left(\beta_k + \frac{1}{\beta_{k-1}}\right) - \left(\lambda + \frac{1}{\lambda}\right)\right\}\hat{v}^{k+1} - \frac{1}{\beta_{k-1}}(\hat{v}^{k+1} - \hat{v}^{k-1}),$$

we deduce from (I) and (II) that $\lim_{k\to\infty} \text{c-lim}_{n\to\infty} \|\vartheta_k\|_n = 0$.

Using these ingredients and the fact that the limiting spectral gap of $A \equiv A(n)$ is strictly positive when $\lambda > 1$, it can be established that

$$\lim_{k\to\infty} \underset{n\to\infty}{\text{c-lim}} \frac{|\langle \hat{v}^k, \hat{\varphi}\rangle_n|}{\|\hat{v}^k\|_n} = 1.$$

This shows that the specific instance (3.28) of the AMP iteration is asymptotically equivalent to the well-known power method for approximating $\hat{\varphi}$, although the dependence of $\hat{v}^0$ on the unknown v means that we cannot use (3.28) as an algorithm in practice. Nevertheless, this asymptotic equivalence ensures that we can apply Theorem 3.1 to obtain the d_2 convergence result in Proposition 3.4 for the joint empirical distribution of the components of $\hat{\varphi}$ and the signal v.

3.4 Confidence intervals and p-values

As a consequence of Theorem 3.1, recall from the discussion after Corollary 3.2 that for fixed k and large n, the AMP iterates (i.e. effective observations) $v^k \equiv v^k(n)$ in the generic procedure (3.3) have the property that $\{(v_i^k - \mu_k v_i)/\sigma_k : 1 \leq i \leq n\}$ behaves approximately like an i.i.d. sample of size n from the $N(0,1)$ distribution. Thus, for a given $\alpha \in [0,1]$, we would expect roughly $n(1-\alpha)$ of these components to have absolute value at most $z_{\alpha/2} := \Phi^{-1}(1-\alpha/2)$, where Φ^{-1} denotes the quantile function of the $N(0,1)$ distribution. Using this observation, we will now outline briefly how to construct confidence intervals for the entries of $v \equiv v(n)$, as well as associated p-values. By Remark 3.3, the (possibly unknown) state evolution parameters μ_k, σ_k can be estimated consistently by $\hat{\mu}_k \equiv \hat{\mu}_k(n) = \bigl(\|v^k\|_n^2 - \|v^{k-1}\|_n^2\bigr)^{1/2}$ and $\hat{\sigma}_k \equiv \hat{\sigma}_k(n) = \|v^{k-1}\|_n$ respectively for each $k \in \mathbb{N}$, so we define

$$\hat{J}_i^k(n,\alpha) := \left[\frac{v_i^k - z_{\alpha/2}\,\hat{\sigma}_k}{\hat{\mu}_k}, \frac{v_i^k + z_{\alpha/2}\,\hat{\sigma}_k}{\hat{\mu}_k}\right] \quad \text{and} \quad p_i^k \equiv p_i^k(n) = 2\left\{1 - \Phi\left(\frac{|v_i^k|}{\hat{\sigma}_k}\right)\right\} \tag{3.29}$$

for $k, n \in \mathbb{N}$, $1 \leq i \leq n$ and $\alpha \in [0,1]$. Montanari and Venkataramanan (2021, Corollary 3.1) showed that for fixed $k \in \mathbb{N}$ and $\alpha \in [0,1]$, the confidence intervals $\hat{J}_1^k(n,\alpha), \ldots, \hat{J}_n^k(n,\alpha)$ have asymptotic mean coverage level $1-\alpha$; specifically,

$$\operatorname*{c-lim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{v_i(n) \in \hat{J}_i^k(n,\alpha)\}} = 1 - \alpha = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbb{P}\bigl(v_i(n) \in \hat{J}_i^k(n,\alpha)\bigr).$$

The first limit above can be established by considering Lipschitz approximations to indicator functions of intervals and appealing to either Theorem 3.1 or 3.5 (for non-spectral and spectral initialisations respectively). The dominated convergence theorem can then be applied to deduce the second equality from the first. Note that for fixed k, α, the asymptotic width of each $\hat{J}_i^k(n,\alpha)$ is $2z_{\alpha/2}/\rho_k$, which is minimised when the empirical Bayes-AMP iterates are used to construct these intervals.

In addition, suppose that the proportion of non-zero entries in the spike $v \equiv v(n)$ tends to $\delta \in (0,1)$ as $n \to \infty$. Then the result cited above asserts that the p-values defined in (3.29) are asymptotically valid for the nulls $\mathcal{N}_n := \{1 \leq i \leq n : v_i \equiv v_i(n) = 0\}$ in the following sense: for any sequence of indices $\bigl(i_0(n) \in \mathcal{N}_n : n \in \mathbb{N}\bigr)$ and all fixed $k \in \mathbb{N}$ and $\alpha \in [0,1]$, we have $\lim_{n\to\infty} \mathbb{P}(p_{i_0(n)}^k \leq \alpha) = \alpha$.

3.5 AMP for more general low-rank matrix estimation problems

Multivariate denoising functions: Thus far, we have only studied estimators $\hat{v}^k = g_k(v^k)$ of the signal v based on a single AMP iterate v^k. One could attempt to improve estimation accuracy by designing estimators $g_k(v^0, \ldots, v^k)$ that also make use of all previous iterates, where $g_k \colon \mathbb{R}^{k+1} \to \mathbb{R}$ is a Lipschitz function that is applied row-wise to $(v^0 \;\cdots\; v^k)$. In the Gaussian spiked model (3.1), consider a more general AMP algorithm with the same initialisers $\hat{v}^{-1}, v^0 \in \mathbb{R}^n$ as in Section 3.1 or 3.2, but with iterates inductively defined by

$$\begin{aligned} \hat{v}^k &:= g_k(v^0, v^1, \ldots, v^k), \quad b_{kj} := \langle \partial_j g_k(v^0, \ldots, v^k)\rangle_n = n^{-1} \textstyle\sum_{i=1}^n \partial_j g_k(v_i^0, \ldots, v_i^k) \;\text{ for } 1 \leq j \leq k \\ v^{k+1} &:= A\hat{v}^k - \textstyle\sum_{j=0}^k b_{kj}\hat{v}^{j-1} \end{aligned} \tag{3.30}$$

for $k \in \mathbb{N}_0$. Here, $\partial_j g_k$ is a weak derivative of the Lipschitz function $(x_0, x_1 \ldots, x_k) \mapsto g_k(x_0, x_1 \ldots, x_k)$ with respect to x_j for $0 \leq j \leq k$. In the setting of Section 3.1, the recursions (3.6)–(3.7) for the state evolution parameters $\mu_k, \sigma_k, \bar{\Sigma}^{[k]}$ instead become

$$\begin{gathered} \mu_{k+1} := \lambda\mathbb{E}\bigl(V g_k(\boldsymbol{\mu_k} V + \bar{\boldsymbol{G}}_{\boldsymbol{k}})\bigr), \qquad \sigma_{k+1}^2 := \mathbb{E}\bigl(g_k(\boldsymbol{\mu_k} V + \bar{\boldsymbol{G}}_{\boldsymbol{k}})^2\bigr), \\ \bar{\Sigma}_{k,\ell} = \bar{\Sigma}_{\ell,k} := \mathbb{E}\bigl(g_{\ell-1}(\boldsymbol{\mu_{\ell-1}} V + \bar{\boldsymbol{G}}_{\boldsymbol{\ell-1}}) \cdot g_{k-1}(\boldsymbol{\mu_{k-1}} V + \bar{\boldsymbol{G}}_{\boldsymbol{k-1}})\bigr) \quad \text{for } 1 \leq \ell \leq k. \end{gathered} \tag{3.31}$$

Here, $\boldsymbol{\mu_\ell} := (\mu_0, \dots, \mu_\ell)$ and $\bar{\boldsymbol{G}}_{\boldsymbol{\ell}} := (\sigma_0 U, \sigma_1 G_1, \dots, \sigma_\ell G_\ell)$ are $(\ell+1)$-dimensional random vectors for $0 \leq \ell \leq k$, with $(\sigma_1 G_1, \dots, \sigma_k G_k) \sim N_k(0, \bar{\Sigma}^{[k]})$ independent of (U, V) from (M1). In the case of spectral initialisation, the recursion (3.15) for $\bar{\Sigma}^{[k]}$ generalises analogously. By relatively straightforward extensions of the proofs of Theorem 2.3 and Corollary 3.2, it follows that

$$\frac{1}{n}\sum_{i=1}^n \psi(\hat{v}_i^k, v_i) \xrightarrow{c} \mathbb{E}\bigl\{\psi\bigl(g_k(\boldsymbol{\mu_k} V + \bar{\boldsymbol{G}}_{\boldsymbol{k}}), V\bigr)\bigr\}, \tag{3.32}$$

provided that the hypotheses of Theorem 3.1 or 3.5 are satisfied. The limiting univariate problem is therefore to estimate $V \sim \pi$ based on $\boldsymbol{\mu_k} V + \bar{\boldsymbol{G}}_{\boldsymbol{k}} = (\mu_0 V + \sigma_0 U, \mu_1 V + \sigma_1 G_1, \dots, \mu_k V + \sigma_k G_k)$.

If the prior distribution π and the initial μ_0, σ_0 are known, then we can recursively define the Bayes-AMP denoisers $\boldsymbol{g}_{\boldsymbol{k}}^* \colon \mathbb{R}^{k+1} \to \mathbb{R}$ and state evolution sequences $\bigl(\mu_k^*, \sigma_k^*, \bar{\Sigma}^{[k]} : k \in \mathbb{N}\bigr)$ via the posterior means

$$\boldsymbol{g}_{\boldsymbol{k}}^*(\boldsymbol{\mu}_{\boldsymbol{k}}^* V + \bar{\boldsymbol{G}}_{\boldsymbol{k}}^*) = \mathbb{E}(V \mid \boldsymbol{\mu}_{\boldsymbol{k}}^* V + \bar{\boldsymbol{G}}_{\boldsymbol{k}}^*) \tag{3.33}$$

and (3.31). Although we might hope that the resulting estimates $\hat{v}^k = \boldsymbol{g}_{\boldsymbol{k}}^*(\boldsymbol{\mu}_{\boldsymbol{k}}^* V + \bar{\boldsymbol{G}}_{\boldsymbol{k}}^*)$ of v are superior to those in Section 3.3, it turns out that in the setting of Theorem 3.10, they are in fact identical to the Bayes-AMP estimates $\hat{v}^{k,\mathrm{B}}$ defined earlier. This is a consequence of the Gaussianity of the noise matrix W in the spiked model, which manifests itself in the following fact (whose proof is given in Section 6.8).

Lemma 3.13. *Let $\bigl(v^0 \equiv v^0(n) : n \in \mathbb{N}\bigr)$ be a sequence of (non-spectral or spectral) initialisers that satisfies either condition (i) or (ii) of Theorem 3.10. Consider the resulting Bayes-AMP iterations $\bigl(v^k \equiv v^k(n) : k, n \in \mathbb{N}\bigr)$ of the form* (3.30) *with $\boldsymbol{g}_{\boldsymbol{k}}^*$ and $\mu_k^*, \sigma_k^*, \bar{\Sigma}^{[k]}$ as above. Then for every $k \in \mathbb{N}$, the random variable $V \sim \pi$ is conditionally independent of $\boldsymbol{\mu}_{\boldsymbol{k-1}}^* V + \bar{\boldsymbol{G}}_{\boldsymbol{k-1}}^*$ given $\mu_k^* V + \sigma_k^* G_k$, whence*

$$\boldsymbol{g}_{\boldsymbol{k}}^*(\boldsymbol{\mu}_{\boldsymbol{k}}^* V + \bar{\boldsymbol{G}}_{\boldsymbol{k}}^*) = \mathbb{E}(V \mid \boldsymbol{\mu}_{\boldsymbol{k}}^* V + \bar{\boldsymbol{G}}_{\boldsymbol{k}}^*) = \mathbb{E}(V \mid \mu_k^* V + \sigma_k^* G_k). \tag{3.34}$$

Consequently, $\boldsymbol{g}_{\boldsymbol{k}}^*$ depends only on its last argument and $b_{kj}^* = 0$ for $1 \leq j \leq k-1$, so $\bigl(v^k \equiv v^k(n) : k, n \in \mathbb{N}\bigr)$ coincides with the sequence $\bigl(v^{k,\mathrm{B}} \equiv v^{k,\mathrm{B}}(n) : k, n \in \mathbb{N}\bigr)$ of Bayes-AMP iterations based on the univariate threshold functions g_k^* in (3.16, 3.17).

Thus, the Bayes-AMP estimates $\hat{v}^{k,\mathrm{B}}$ in Section 3.3 are actually Bayes optimal over a wider class of estimators arising from AMP algorithms (3.30) with separable multivariate denoisers. In particular, to define the Bayes-AMP denoising functions in this Gaussian setting, it suffices to track the state evolution scalars μ_k^*, σ_k^* rather than the full limiting covariance matrices $\bar{\Sigma}^{[k]}$. As we will see below, this is in sharp contrast to AMP iterations with general non-Gaussian rotationally invariant matrices. Note also that the conditional independence in Lemma 3.13 is specific to Bayes-AMP, and is not guaranteed to hold for the state evolution sequence and limiting random variables associated with a general AMP iteration of the form (3.30).

Estimation of a rectangular rank-one matrix: Let $A \in \mathbb{R}^{n \times p}$ be an observation matrix given by

$$A \equiv A(n) = \frac{\lambda}{n} u v^\top + W', \tag{3.35}$$

where W' is a Gaussian noise matrix with $W'_{ij} \overset{\mathrm{iid}}{\sim} N(0, 1/n)$ for $1 \leq i \leq n$ and $1 \leq j \leq p$, and seek to estimate one or both of the unknown vectors $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^p$.

An important example of this observation scheme is a spiked covariance model (Johnstone, 2006; Johnstone and Lu, 2009) where $a_1, \dots, a_n \overset{\mathrm{iid}}{\sim} N_p(0, \Sigma)$ with $\Sigma := (\lambda v v^\top + I_p)/n \in \mathbb{R}^{p \times p}$. In this case, the matrix $A \in \mathbb{R}^{n \times p}$ with rows $a_1, \dots, a_n$ is of the form (3.35) with $u \sim N_n(0, I_n)$.

By analogy with the symmetric case in Section 3.1, an AMP algorithm for the model (3.35) can be obtained by replacing the Gaussian matrix W in the abstract asymmetric AMP iteration (2.10) with the data matrix A (Rangan and Fletcher, 2012; Deshpande and Montanari, 2014). For $k \in \mathbb{N}_0$ and generic sequences of Lipschitz functions $(f_k)_{k=0}^{\infty}$ and $(g_k)_{k=0}^{\infty}$ satisfying (M2), the corresponding AMP procedure takes the form

$$\begin{aligned} u^k &:= Af_k(v^k) - b_k g_{k-1}(u^{k-1}), & c_k &:= n^{-1}\textstyle\sum_{i=1}^n g_k'(u_i^k), \\ v^{k+1} &:= A^\top g_k(u^k) - c_k f_k(v^k), & b_{k+1} &:= n^{-1}\textstyle\sum_{i=1}^p f_{k+1}'(v_i^{k+1}) \end{aligned} \tag{3.36}$$

for $k \in \mathbb{N}_0$. Based on an appropriate state evolution recursion, analogues of Theorems 3.1 and 3.5 can be formulated for (3.36) with non-spectral and spectral initialisations respectively. These results apply to an asymptotic regime where $n, p \to \infty$ with $n/p \to \delta$ for some $\delta \in (0,1)$, and where a version of (M1) holds (with $\|u\|_n, \|v\|_p \xrightarrow{c} 1$ and the empirical distributions of the components of u and v converging completely in d_2 to suitable limits). A suitable spectral initialiser for (3.36) is $v^0 = \hat{\varphi}^R$, a principal right singular vector of A with $\|\hat{\varphi}^R\|_p = 1$ (Montanari and Venkataramanan, 2021, Section 4). The associated spectral threshold is at $\sqrt{\delta}$: if $n/p \to \delta$ and $\lambda > 1/\sqrt{\delta}$, then the limiting empirical correlation $|\langle \hat{\varphi}^R, v\rangle_n|/\|v\|_n$ is strictly positive (Paul, 2007; Bai and Silverstein, 2010).

Estimation of rank-s matrices for $s > 1$: The general rank-s spiked models take the form

$$A = \sum_{j=1}^s \frac{\lambda_j}{n} v_j v_j^\top + W \quad \text{(symmetric)}; \qquad A = \sum_{j=1}^s \frac{\lambda_j}{n} v_j u_j^\top + W' \quad \text{(asymmetric)} \tag{3.37}$$

for some $\lambda_1 \geq \cdots \geq \lambda_s$, where $(v_j \in \mathbb{R}^n : 1 \leq j \leq s)$ and $(u_j \in \mathbb{R}^p : 1 \leq j \leq s)$ are sets of unknown orthogonal vectors (with $\|v_j\|_n, \|u_j\|_p \xrightarrow{c} 1$ for all j), and where the Gaussian noise matrices W, W' are as in (3.3) and (3.36) respectively. Parker et al. (2014a,b), Kabashima et al. (2016), Lesieur et al. (2017) and Montanari and Venkataramanan (2021) proposed generalisations of the AMP algorithms (3.3) and (3.36) for estimating $u_1, \ldots, u_s, v_1, \ldots, v_s$ (and hence the signal matrices) in (3.37). For $s > 1$, the main difference with the rank-one case is that the iterates in these procedures are matrices rather than vectors.

More precisely, in the symmetric case, each iterate is an $n \times s$ matrix to which a row-wise thresholding function is applied to obtain updated estimates of $v_1, \ldots, v_s$. When the initialiser is a matrix consisting of eigenvectors corresponding to the s largest eigenvalues of A, a rigorous state evolution result was obtained by Montanari and Venkataramanan (2021, Section 6). Similarly, in the rectangular case, the iterates are $n \times s$ and $p \times s$ matrices and we can take the columns of the initialising matrix to be right singular vectors of A. Additional complications arise when $\lambda_1, \ldots, \lambda_s$ are not pairwise distinct. In these degenerate cases, not all of the vectors $v_1, \ldots, v_s$ are identifiable up to sign in (3.37). Indeed, if some λ occurs with multiplicity $r > 1$, then by applying any orthogonal transformation that fixes the r-dimensional subspace spanned by the corresponding signal vectors, we can obtain another valid representation of the signal matrix while leaving A unchanged. There is therefore an inherent ambiguity in the definition of the spectral initialisers above. By analogy with Remark 3.6 for the $s = 1$ case, the AMP state evolution consequently has a random initialiser that is defined using a Haar-distributed orthogonal matrix. Conditioned on this initialisation, the state evolution parameters for the subsequent iterations are deterministic.

Universality: As mentioned in the Introduction, the theoretical framework for AMP was originally built around Gaussian random matrices, but the conclusions of Theorems 3.1 and 3.5 (as well as the master theorems in Section 2) have now been extended to encompass more general random matrix ensembles. In so-called 'spiked Wigner' models of the form (3.1), the symmetric noise matrices $W \equiv W(n)$ have independent upper-triangular entries $(W_{ij} : 1 \leq i \leq j \leq n)$ that are uniformly subexponential across $n \in \mathbb{N}$ with $\mathbb{E}(W_{ij}) = 0$ and $\mathrm{Var}(W_{ij}) = (1 + \delta_{ij})/n$. It was previously known that the eigenstructure of the corresponding observation matrix A undergoes the BBP phase transition described in Section 3.1 at the same spectral threshold $\lambda = 1$ as for 'spiked GOE' matrices; see for instance Anderson et al. (2010), Knowles and Yin (2013) and Perry et al. (2018). Recently, Chen and Lam (2021, Examples 2.1 and 2.2) used the method of Slepian interpolation to prove that in AMP algorithms of the form (3.30) based on matrices A from rank-one spiked Wigner models, the iterates have the same asymptotics as in the original Gaussian setting, with or without spectral initialisation.

In a different direction, Fan (2022) developed a more general class of AMP procedures for symmetric and rectangular rank-one spiked models (3.1, 3.35) in which the noise matrices are orthogonally invariant. In the symmetric case, this means that $W \equiv W(n)$ satisfies $W \overset{d}{=} Q^\top W Q$ for all deterministic orthogonal $Q \in \mathbb{R}^{n\times n}$, and it can be shown that the only such W with independent, mean-zero upper-triangular entries are scalar multiples of GOE matrices (e.g. Mehta, 2004). For other orthogonally invariant W, Opper et al. (2016) and Fan (2022, Section 3) modified the AMP algorithm (3.30) to allow v^{k+1} to depend on all previous iterates via

$$\hat{v}^k = g_k(v^0, v^1, \dots, v^k) \quad \text{and} \quad v^{k+1} = A\hat{v}^k - \sum_{j=0}^{k} b_{kj}\hat{v}^{j-1} \quad \text{for } k \in \mathbb{N}_0, \tag{3.38}$$

in a such a way that the joint empirical distributions have well-defined Wasserstein limits. To achieve this, the technical crux is to design suitable Onsager coefficients $b_{k1}, \dots, b_{kk}$ that depend on the limiting spectral distribution of W (when it exists) through its moments and free cumulants, which also determine the corresponding state evolution sequences. Asymptotic convergence results similar in spirit to Theorems 3.1 and 3.10 can then be established for iterations of the form (3.38) and their Bayes-AMP versions. In particular, for each k, the components of the effective observation v^k behave like those of $\mu_k v + \sigma_k \xi$ for large n, where $\xi \sim N_n(0, I_n)$ is independent of v as before but the recursions for μ_k, σ_k are significantly more involved than (3.31). As in Section 3.3, it turns out that for large k, the Bayes-AMP estimates $\hat{v}^k$ of the spike v can substantially improve on the spectral estimator (namely a leading eigenvector of the observation matrix A) in terms of asymptotic mean squared error (Fan, 2022, Remark 3.2). Unlike in the Gaussian setting of Lemma 3.13 however, the Bayes optimal denoisers $\boldsymbol{g}_k^*\colon \mathbb{R}^{k+1} \to \mathbb{R}$ may depend on all of their arguments and must be defined with respect to the full limiting covariance structure (Fan, 2022, Remark 3.3).

4 GAMP for generalised linear models

In this section, we give a unified treatment of a class of AMP algorithms for models of the following generic form: suppose that we generate a design matrix $X \in \mathbb{R}^{n\times p}$ with rows $x_1, \dots, x_n \in \mathbb{R}^p$, and observe $y \equiv (y_1, \dots, y_n) \in \mathbb{R}^n$ satisfying

$$y_i = h(x_i^\top \beta, \varepsilon_i) \quad \text{for } i = 1, \dots, n, \tag{4.1}$$

where $\beta \equiv (\beta_1, \dots, \beta_p)$ is the target of inference, $\varepsilon \equiv (\varepsilon_1, \dots, \varepsilon_n)$ is a vector of noise variables and $h\colon \mathbb{R}^2 \to \mathbb{R}$ is a known function. We will focus on the random design setting where $x_1, \dots, x_n \overset{\text{iid}}{\sim} N_p(0, I_p/n)$, which is a common assumption in high-dimensional statistics and compressed sensing. Frequently, $\varepsilon_1, \dots, \varepsilon_n$ are assumed to be independent of each other and of X, in which case (4.1) becomes

$$y_i \,|\, x_i \sim Q_i(\cdot \,|\, x_i^\top \beta), \tag{4.2}$$

where $Q_i(\cdot \,|\, z)$ denotes the distribution of $h(z, \varepsilon_i)$ for a fixed $z \in \mathbb{R}$ and $1 \le i \le n$. In statistics, (4.2) is traditionally referred to as a *generalised linear model* (GLM) for $(x_1, y_1), \dots, (x_n, y_n)$ if the conditional distributions of y_i given x_i have densities of *exponential dispersion family form* (Pace and Salvan, 1997)

$$u \mapsto a(\sigma_i^2, u) \exp\left\{\frac{u\Theta(\mu_i) - K(\Theta(\mu_i))}{\sigma_i^2}\right\} \tag{4.3}$$

with respect to either Lebesgue measure on $\mathbb{R}$ or counting measure on $\mathbb{Q}$. In (4.3), the mean parameter $\mu_i \in \mathcal{M} \subseteq \mathbb{R}$ is related to x_i via $\mu_i = \eta^{-1}(x_i^\top \beta)$ for some strictly increasing, twice differentiable *link function* η, and $\sigma_i \in \mathcal{D} \subseteq (0, \infty)$ is the *dispersion parameter*, while a, K, Θ are fixed functions with $K'' > 0$ on $\mathbb{R}$ and $\Theta = (K')^{-1}$. The GLM framework encompasses a broad class of parametric models, including the standard linear model, phase retrieval (where $y_i = (x_i^\top \beta)^2 + \varepsilon_i$ for $1 \le i \le n$), and logistic, binomial and Poisson regression (e.g. McCullagh and Nelder, 1989; Agresti, 2015). Sometimes, 'GLM' is used as an umbrella term to describe more general models of the form (4.1, 4.2).

Likelihood-based inference for β in (4.2, 4.3) is justified by classical asymptotic theory when p is fixed and $n \to \infty$, or when p grows sufficiently slowly with n (Portnoy, 1984, 1985, 1988). However, in modern high-dimensional regimes where $n, p \to \infty$ and the aspect ratio n/p of the design matrix X is bounded, different tools are needed to construct and analyse estimators of β, and it is in this context that we introduce the GAMP paradigm below.

4.1 Master theorem for GAMP

The *generalised AMP* (GAMP) *algorithm* proposed by Rangan (2011) iteratively produces estimates $\hat{\beta}^k, \theta^k$ of $\beta \in \mathbb{R}^p$ and $\theta := X\beta \in \mathbb{R}^n$ respectively in (4.1), via update steps of the following form: given $\hat{r}^{-1} := 0 \in \mathbb{R}^n$, $b_0 \in \mathbb{R}$ and an initialiser $\hat{\beta}^0 \in \mathbb{R}^p$, recursively define

$$\begin{aligned} \theta^k &:= X\hat{\beta}^k - b_k\hat{r}^{k-1}, & \hat{r}^k &:= g_k(\theta^k, y), & c_k &:= n^{-1}\textstyle\sum_{i=1}^n g_k'(\theta_i^k, y_i), \\ \beta^{k+1} &:= X^\top \hat{r}^k - c_k\hat{\beta}^k, & \hat{\beta}^{k+1} &:= f_{k+1}(\beta^{k+1}), & b_{k+1} &:= n^{-1}\textstyle\sum_{j=1}^p f_{k+1}'(\beta_j^{k+1}) \end{aligned} \tag{4.4}$$

for $k \in \mathbb{N}_0$. Here, $g_k : \mathbb{R}^2 \to \mathbb{R}$ and $f_{k+1} : \mathbb{R} \to \mathbb{R}$ are Lipschitz in their first argument, and $g_k' : \mathbb{R}^2 \to \mathbb{R}$, $f_{k+1}' : \mathbb{R} \to \mathbb{R}$ agree with the partial derivatives of g_k, f_{k+1} respectively with respect to their first arguments, wherever the latter are defined. As in previous sections, these functions are understood to act componentwise on their vector arguments in (4.4). The goal of Section 4 is to develop the theory and applications of GAMP, whose statistical utility can be summarised in the following key points:

(i) *Exact asymptotic characterisation via state evolution*: The Onsager correction terms $-b_k\hat{r}^{k-1}$, $-c_k\hat{\beta}^k$ are designed to ensure that in a high-dimensional limiting regime where $n, p \to \infty$ with $n/p \to \delta \in (0, \infty)$, the empirical distributions of the entries of the iterates in (4.4) converge to well-defined Wasserstein limits. These asymptotic distributions are characterised by the state evolution recursion (4.6)–(4.7) below. Consequently, for each fixed $k \in \mathbb{N}_0$, the entries of $\hat{\beta}^{k+1} \in \mathbb{R}^p$ have approximately the same empirical distribution as those of $f_{k+1}(\mu_k\beta + \sigma_k\xi)$ when p is large; here, $\beta \in \mathbb{R}^p$ is the unknown signal, $\xi \sim N_p(0, I_p)$ is an independent noise vector, μ_k, σ_k are the effective signal strength and noise level respectively, and f_{k+1} can be viewed as a denoising function. This result facilitates a targeted approach to inference for structured signals β, whereby informed choices of $(f_k, g_k : k \in \mathbb{N}_0)$ can be made to accommodate different types of prior information (Section 4.2).

(ii) *Link to convex optimisation problems*: For suitable choices of f_k, g_k, the GAMP recursion (4.4) can be interpreted as an alternating minimisation procedure for solving a convex optimisation problem of the form (4.22), and the fixed points of this iteration are minimisers of the convex objective function (Proposition 4.4 in Section 4.4). Together with the state evolution description of (4.4), this forms the basis of a *systematic* approach to deriving exact performance guarantees for the Lasso and other (penalised or unpenalised) M-estimators in high-dimensional GLMs (Sections 4.5–4.7).

In this subsection, we address point (i) above and formally state a 'master theorem' for GAMP (Theorem 4.2). Consider a sequence of recursions (4.4) indexed by $n \in \mathbb{N}$ and $p \equiv p_n$, where $n/p \to \delta \in (0, \infty)$ as $n \to \infty$, and assume that

(G0) For each n, the design matrix $X \equiv X(n) \in \mathbb{R}^{n \times p}$ has i.i.d. $N(0, 1/n)$ entries and is independent of $\big(\hat{\beta}^0(n), \beta(n), \varepsilon(n)\big) \in \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^n$.

At first sight, it would appear that the GAMP algorithm (4.4) is an instance of the abstract asymmetric AMP recursion (2.10), but in models (4.1) where (G0) holds, the crucial difference in the probabilistic structure is that the observation vector $y \equiv y(n) \in \mathbb{R}^n$ is in general not independent of $X \equiv X(n)$. This means that condition (B0) does not hold with $\gamma = y$, so the original master theorem for asymmetric AMP (Theorem 2.5) cannot be directly applied in this setting, and in fact does not give the correct limiting distributions for (4.4).

Instead, Theorem 4.2 below is derived from a general state evolution result for *matrix-valued* AMP iterations (Section 6.7), under suitable analogues of (B1)–(B5) on the inputs to the GAMP recursions (4.4) as $n \to \infty$ and $n/p \to \delta$: for some $r \in [2, \infty)$, suppose that

(G1) There exist random variables $\bar{\beta} \sim \pi_{\bar{\beta}}$ and $\bar{\varepsilon} \sim P_{\bar{\varepsilon}}$ with $\mathbb{E}(\bar{\beta}^2) > 0$ and $\mathbb{E}(|\bar{\beta}|^r), \mathbb{E}(|\bar{\varepsilon}|^r) < \infty$, such that writing $\nu_p(\beta)$ and $\nu_n(\varepsilon)$ for the empirical distributions of the components of $\beta \equiv \beta(p)$ and $\varepsilon \equiv \varepsilon(n)$ respectively, we have $d_r\big(\nu_p(\beta), \pi_{\bar{\beta}}\big) \xrightarrow{c} 0$ and $d_r\big(\nu_n(\varepsilon), P_{\bar{\varepsilon}}\big) \xrightarrow{c} 0$.

(G2) $\|\hat{\beta}^0\|_{p,r} = O_c(1)$ and there exists a non-negative definite $\Sigma_0 \in \mathbb{R}^{2\times 2}$ such that $\breve{\beta}^0 := (\beta \ \hat{\beta}^0) \in \mathbb{R}^{p\times 2}$ satisfies

$$\frac{1}{n}(\breve{\beta}^0)^\top \breve{\beta}^0 = \frac{1}{n}\begin{pmatrix} \beta^\top\beta & \beta^\top\hat{\beta}^0 \\ (\hat{\beta}^0)^\top\beta & (\hat{\beta}^0)^\top\hat{\beta}^0 \end{pmatrix} \xrightarrow{c} \Sigma_0.$$

(G3) There exists a Lipschitz $F_0 \colon \mathbb{R} \to \mathbb{R}$ such that $\langle \hat{\beta}^0, \phi(\beta)\rangle_p \xrightarrow{c} \mathbb{E}\big(F_0(\bar{\beta})\phi(\bar{\beta})\big)$ and $\mathbb{E}\big(F_0(\bar{\beta})^2\big) \leq (\Sigma_0)_{22}$ for all Lipschitz $\phi \colon \mathbb{R} \to \mathbb{R}$.

(G4) For each $k \in \mathbb{N}_0$, the function f_{k+1} is non-constant on $\mathbb{R}$, and $\tilde{g}_k \colon (z, u, v) \mapsto g_k(u, h(z, v))$ is Lipschitz on $\mathbb{R}^3$ with $P_{\bar{\varepsilon}}\big(\{v : (z, u) \mapsto \tilde{g}_k(z, u, v) \text{ is non-constant}\}\big) > 0$.

We remark here that while (G2) is in general a stronger requirement than (B2), both (G2) and (G3) are implied by (G1) if for some fixed $c \in \mathbb{R}$ we have $\hat{\beta}^0 \equiv \hat{\beta}^0(n) = c\mathbf{1}_p$ for all n. As in Section 3, constraints on $\beta \equiv \beta(n)$ such as sparsity or entrywise non-negativity will be reflected in the form of the 'limiting prior distribution' $\pi_{\bar{\beta}}$. Note that

$$(\Sigma_0)_{11} = \operatorname*{c-lim}_{n\to\infty}\left(\frac{p}{n} \cdot \frac{\|\beta\|^2}{p}\right) = \frac{\mathbb{E}(\bar{\beta}^2)}{\delta} > 0 \tag{4.5}$$

by (G1) and (G2). Also, the condition on $\varepsilon \equiv \varepsilon(n)$ in (G1) is satisfied if $\varepsilon_1, \dots, \varepsilon_n \overset{\text{iid}}{\sim} P_{\bar{\varepsilon}}$ for each n.

State evolution: With Σ_0 as in (G2), the state evolution parameters $\big(\mu_k \in \mathbb{R},\ \sigma_k \in [0, \infty),\ \Sigma_k \in \mathbb{R}^{2\times 2} : k \in \mathbb{N}\big)$ are recursively defined by

$$\mu_{k+1} := \mathbb{E}\big(\partial_z \tilde{g}_k(Z, Z_k, \bar{\varepsilon})\big), \qquad \sigma_{k+1}^2 := \mathbb{E}\big(\tilde{g}_k(Z, Z_k, \bar{\varepsilon})^2\big) = \mathbb{E}\big(g_k(Z_k, Y)^2\big), \tag{4.6}$$

$$\Sigma_{k+1} := \frac{1}{\delta}\begin{pmatrix} \mathbb{E}(\bar{\beta}^2) & \mathbb{E}\{\bar{\beta} f_{k+1}(\mu_{k+1}\bar{\beta} + \sigma_{k+1}G_{k+1})\} \\ \mathbb{E}\{\bar{\beta} f_{k+1}(\mu_{k+1}\bar{\beta} + \sigma_{k+1}G_{k+1})\} & \mathbb{E}\{f_{k+1}(\mu_{k+1}\bar{\beta} + \sigma_{k+1}G_{k+1})^2\} \end{pmatrix} \tag{4.7}$$

for $k \in \mathbb{N}_0$, where we take $(Z, Z_k) \sim N_2(0, \Sigma_k)$ to be independent of $\bar{\varepsilon} \sim P_{\bar{\varepsilon}}$, define $Y := h(Z, \bar{\varepsilon})$, and take $G_{k+1} \sim N(0, 1)$ to be independent of $\bar{\beta} \sim \pi_{\bar{\beta}}$. Under (G4), it can be shown as in Lemma 2.2 that if $\sigma_1 > 0$, then $\sigma_k > 0$ and Σ_k is positive definite for all $k \in \mathbb{N}$. In (4.6), $\partial_z \tilde{g}_k$ denotes the partial derivative of $\tilde{g}_k$ with respect to its first argument; observe that by (G4), $z \mapsto \tilde{g}_k(z, u, v)$ is Lipschitz and hence differentiable almost everywhere for all $(u, v) \in \mathbb{R}^2$, so μ_{k+1} is well-defined.

Stein's lemma (Lemma 6.20) can be used to derive some alternative expressions for μ_{k+1} that will be useful later on; see Mondelli and Venkataramanan (2020, Proposition 3.1) or Section 6.9 for the proof of the following lemma.

Lemma 4.1. *For each $k \in \mathbb{N}$, letting $\tilde{G}_k \sim N(0, 1)$ be independent of $(Z, \bar{\varepsilon})$, we have $(Z, Z_k, \bar{\varepsilon}) \overset{d}{=} (Z, \mu_{Z,k}Z + \sigma_{Z,k}\tilde{G}_k, \bar{\varepsilon})$, where*

$$\begin{aligned} \mu_{Z,k} &:= \frac{\mathbb{E}\big(\bar{\beta} f_k(\mu_k\bar{\beta} + \sigma_k\tilde{G}_k)\big)}{\mathbb{E}(\bar{\beta}^2)} = \frac{\Sigma_{21}}{\Sigma_{11}}, \\ \sigma_{Z,k}^2 &:= \frac{\mathbb{E}(\bar{\beta}^2)\,\mathbb{E}\big(f_k(\mu_k\bar{\beta} + \sigma_k\tilde{G}_k)^2\big) - \mathbb{E}\big(\bar{\beta} f_k(\mu_k\bar{\beta} + \sigma_k\tilde{G}_k)\big)^2}{\delta\,\mathbb{E}(\bar{\beta}^2)} = \Sigma_{22} - \frac{\Sigma_{12}^2}{\Sigma_{11}}, \end{aligned} \tag{4.8}$$

with $\Sigma \equiv \Sigma_k$. Thus, $\mu_{k+1} = \mathbb{E}\big(\partial_z \tilde{g}_k(Z, \mu_{Z,k}Z + \sigma_{Z,k}\tilde{G}_k, \bar{\varepsilon})\big)$ and $\sigma_{k+1}^2 = \mathbb{E}\big(\tilde{g}_k(Z, \mu_{Z,k}Z + \sigma_{Z,k}\tilde{G}_k, \bar{\varepsilon})^2\big)$. Moreover,

$$\mu_{k+1} = \frac{\delta}{\mathbb{E}(\bar{\beta}^2)}\mathbb{E}\big(Zg_k(Z_k, Y)\big) - \mu_{Z,k}\,\mathbb{E}\big(g_k'(Z_k, Y)\big) = \mathbb{E}\left(\frac{\mathbb{E}(Z \mid Z_k, Y) - \mathbb{E}(Z \mid Z_k)}{\operatorname{Var}(Z \mid Z_k)}\, g_k(Z_k, Y)\right). \tag{4.9}$$

Before stating the main result of this subsection, we make an further regularity assumption that is similar to (B5).

(G5) For each $k \in \mathbb{N}_0$, writing $D_k \subseteq \mathbb{R}^2$ for the set of discontinuities of g_k', we have $\mathbb{P}\big((Z_k, Y) \in D_k\big) = 0$, and f_{k+1}' is continuous Lebesgue almost everywhere.

Theorem 4.2. *Suppose that (G0)–(G5) hold for a sequence of GAMP recursions (4.4) indexed by n and $p \equiv p_n$, with $n/p \to \delta \in (0,\infty)$ and $\sigma_1 > 0$. Then for each $k \in \mathbb{N}_0$, we have*

$$\sup_{\psi \in \mathrm{PL}_2(r,1)} \left| \frac{1}{p}\sum_{j=1}^{p} \psi(\beta_j^{k+1}, \beta_j) - \mathbb{E}\big(\psi(\mu_{k+1}\bar{\beta} + \sigma_{k+1}G_{k+1}, \bar{\beta})\big)\right| \xrightarrow{c} 0, \tag{4.10}$$

$$\sup_{\psi \in \mathrm{PL}_3(r,1)} \left| \frac{1}{n}\sum_{i=1}^{n} \psi(\theta_i^{k}, \theta_i, \varepsilon_i) - \mathbb{E}\big(\psi(\mu_{Z,k}Z + \sigma_{Z,k}\tilde{G}_k, Z, \bar{\varepsilon})\big)\right| \xrightarrow{c} 0 \tag{4.11}$$

as $n, p \to \infty$ with $n/p \to \delta$, where $\theta_i \equiv \theta_i(n) = x_i^\top \beta$ for $n \in \mathbb{N}$ and $1 \le i \le n$.

Writing $\nu_p(\beta^k, \beta)$ for the joint empirical distribution of the components of $\beta^k, \beta \in \mathbb{R}^p$, and $\breve{\nu}^k$ for the distribution of $(\mu_k\bar{\beta} + \sigma_k G_k, \bar{\beta})$, we can express the conclusion of (4.10) as

$$\tilde{d}_r\big(\nu_p(\beta^k, \beta), \breve{\nu}^k\big) \xrightarrow{c} 0, \quad \text{or equivalently} \quad d_r\big(\nu_p(\beta^k, \beta), \breve{\nu}^k\big) \xrightarrow{c} 0 \quad \text{as } n \to \infty.$$

Likewise, (4.11) says that the joint empirical distribution $\nu_n(\theta^k, \theta, \varepsilon)$ converges completely in d_r to the distribution of $(\mu_{Z,k}Z + \sigma_{Z,k}\tilde{G}_k, Z, \bar{\varepsilon}) \overset{d}{=} (Z_k, Z, \bar{\varepsilon})$.

Interpretation: Informally, when p is large, the components of β^k have approximately the same empirical distribution as those of $\mu_k\beta + \sigma_k\xi$, where $\xi \sim N_p(0, I_p)$ is independent of $\beta \in \mathbb{R}^p$. By analogy with the limiting univariate problem of estimating $\bar{\beta} \sim \pi_{\bar{\beta}}$ based on a corrupted observation $\mu_k\bar{\beta} + \sigma_k G_k$, we can regard β^k as an *effective observation* and $\rho_k := (\mu_k/\sigma_k)$ as an *effective signal-to-noise ratio*; recall the discussion after Corollary 3.2.

Remark 4.3. Similarly to Remark 2.4, it turns out that in the setting of Theorem 4.2, condition (G5) ensures that

$$\begin{aligned} c_k &= \frac{1}{n}\sum_{i=1}^{n} g_k'(\theta_i^k, y_i) \xrightarrow{c} \mathbb{E}\big(g_k'(Z_k, Y)\big) =: \bar{c}_k, \\ b_{k+1} &= \frac{1}{n}\sum_{j=1}^{p} f_{k+1}'(\beta_j^{k+1}) \xrightarrow{c} \frac{\mathbb{E}\big(f_{k+1}'(\mu_{k+1}\bar{\beta} + \sigma_{k+1}G_{k+1})\big)}{\delta} =: \bar{b}_{k+1} \end{aligned} \tag{4.12}$$

as $n, p \to \infty$ with $n/p \to \delta$, for each $k \in \mathbb{N}_0$. In fact, the theorem holds under (G0)–(G4) if b_k, c_k are replaced with $\bar{b}_k, \bar{c}_k$ respectively in (4.4), in which case (G5) is not needed.

By defining an augmented state evolution that specifies the covariance structure of the limiting Gaussians $G_1, G_2, \dots$ and $\tilde{G}_1, \tilde{G}_2, \dots$, we can establish the d_r limits of the joint empirical distributions $\nu_p(\beta^1, \dots, \beta^k, \beta)$ and $\nu_n(\theta^0, \dots, \theta^k, \theta)$, similarly to (3.7) and Theorem 3.1. For simplicity of presentation, we do not state this stronger conclusion. Its proof is identical in most respects to that of Theorem 4.2, which we now summarise.

Proof (sketch) of Theorem 4.2. As mentioned previously, the overall objective is to handle the dependence of y on X (through $\theta = X\beta$) in (4.4), and show that the 'noise' component $\tilde{\beta}^k \equiv \tilde{\beta}^k(n) := \beta^k - \mu_k\beta$ of the effective observations is approximately Gaussian (and independent of β) for large n. To this end, consider rewriting the second update step as

$$\tilde{\beta}^{k+1} \equiv \beta^{k+1} - \mu_{k+1}\beta = X^\top \tilde{g}_k(\theta, \theta^k, \varepsilon) - \begin{pmatrix} \beta & f_k(\tilde{\beta}^k + \mu_k\beta) \end{pmatrix} \begin{pmatrix} \mu_{k+1} \\ \langle \partial_2 \tilde{g}_k(\theta, \theta^k, \varepsilon)\rangle_n \end{pmatrix}. \tag{4.13}$$

Here, $\tilde{g}_k(\theta, \theta^k, \varepsilon) = g_k(\theta^k, y) = \hat{r}^k$ (applying $\tilde{g}_k$ componentwise), $f_k(\tilde{\beta}^k + \mu_k\beta) = f_k(\beta^k) = \hat{\beta}^k$ and $\partial_2\tilde{g}_k(z, u, v) := g_k'(u, h(z, v))$ agrees with the partial derivative of $\tilde{g}_k$ with respect to its second argument, wherever the latter is defined. A useful feature of (4.13) is that unlike y, the noise vector ε is independent of X by (G0). Since both θ and θ^k depend on X, this suggests treating $\tilde{\theta}^k := (\theta\;\theta^k) \in \mathbb{R}^{n\times 2}$ as a single entity, and rewriting the first update step in (4.4) as

$$\tilde{\theta}^k \equiv (\theta\;\;\theta^k) = X\left(\beta\quad f_k(\tilde{\beta}^k + \mu_k\beta)\right) - \tilde{g}_{k-1}(\theta, \theta^{k-1}, \varepsilon)\left(0\quad \frac{p}{n}\langle f_k'(\tilde{\beta}^k + \mu_k\beta)\rangle_p\right), \tag{4.14}$$

where $\frac{p}{n}\langle f_k'(\tilde{\beta}^k + \mu_k\beta)\rangle_p = \frac{p}{n}\langle f_k'(\beta^k)\rangle_p = b_k$. In doing so, we have recast (4.4) as a matrix-valued AMP iteration (4.13)–(4.14) that is no longer a valid algorithm for practical purposes, but is more amenable to theoretical analysis. Indeed, its asymptotics can be derived by applying a master theorem for abstract recursions (6.52) of this type; see Section 6.7. The significance of the definition of $\mu_{k+1} = \mathbb{E}\big(\partial_z\tilde{g}_k(Z, Z_k, \bar{\varepsilon})\big) \equiv \mathbb{E}\big(\partial_1\tilde{g}_k(Z, Z_k, \bar{\varepsilon})\big)$ in (4.6) is that the final term in (4.13) is a non-linear correction based on the derivative (gradient) of $\tilde{g}_k$. The final term in (4.14) has a similar interpretation as a multivariate analogue of the original b_k in (4.4), and together these ensure that the limiting empirical distributions of the iterates in (4.13)–(4.14) are indeed Gaussian. □

4.2 Choosing the functions f_k, g_k, and inference for β

Asymptotic estimation error: Since the functions f_k in (4.4) are Lipschitz by assumption, it follows as in Corollary 3.2 that in the setting of Theorem 4.2 above, the asymptotic estimation error of $\hat{\beta}^k$ with respect to any loss function $\psi \in \mathrm{PL}_2(r)$ is given by

$$\frac{1}{p}\sum_{j=1}^{p}\psi(\hat{\beta}_j^k, \beta_j) \xrightarrow{c} \mathbb{E}\big\{\psi\big(f_k(\mu_k\bar{\beta} + \sigma_k G_k), \bar{\beta}\big)\big\} \tag{4.15}$$

for each $k \in \mathbb{N}$, as $n, p \to \infty$ with $n/p \to \delta$. In particular, taking $\psi(x, y) = |x - y|^q$ for $q \in [1, r]$, we obtain the asymptotic normalised ℓ_q error $\text{c-lim}_{p\to\infty}\, p^{-1}\|\hat{\beta}^k - \beta\|_q = \mathbb{E}\big\{\big(f_k(\mu_k\bar{\beta} + \sigma_k G_k) - \bar{\beta}\big)^q\big\}^{1/q}$.

Bayes-GAMP: If the limiting prior distribution $\pi_{\bar{\beta}}$, the limiting noise distribution $P_{\bar{\varepsilon}}$ and the initial $\Sigma_0 \in \mathbb{R}^{2\times 2}$ are known, then guided by Lemma 3.7, we can proceed as in Section 3.3 and choose f_k, g_k in (4.4) so as to maximise the effective signal-to-noise ratios $\rho_k = (\mu_k/\sigma_k)^2$ and $\rho_{Z,k} := (\mu_{Z,k}/\sigma_{Z,k})^2$ for each k.

Specifically, given the matrix $\Sigma \equiv \Sigma_k \in \mathbb{R}^{2\times 2}$ in (4.7) for some $k \in \mathbb{N}_0$, we can obtain $\mu_{Z,k}, \sigma_{Z,k}$ from (4.8); conversely, given $\mu_{Z,k}, \sigma_{Z,k}$, we can recover Σ since $\Sigma_{11} = \delta^{-1}\mathbb{E}(\bar{\beta}^2)$ is known, and (4.8) yields $\Sigma_{21} = \Sigma_{11}\mu_{Z,k}$ and $\Sigma_{22} = \sigma_{Z,k}^2 + \Sigma_{11}\mu_{Z,k}^2$. Now take $(Z, Z_k) \sim N_2(0, \Sigma_k)$ to be independent of $\bar{\varepsilon} \sim P_{\bar{\varepsilon}}$, and let $Y = h(Z, \bar{\varepsilon})$, so that Y and Z_k are conditionally independent given Z. Based on the joint distribution of (Z, Z_k, Y), let $g_k^*\colon \mathbb{R}^2 \to \mathbb{R}$ be a measurable function satisfying

$$g_k^*(Z_k, Y) = \frac{\mathbb{E}(Z \,|\, Z_k, Y) - \mathbb{E}(Z \,|\, Z_k)}{\mathrm{Var}(Z \,|\, Z_k)}, \tag{4.16}$$

where $\mathbb{E}(Z \,|\, Z_k) = m_k Z_k$ with

$$m_k := \frac{\Sigma_{21}}{\Sigma_{22}} = \frac{\mu_{Z,k}}{\sigma_{Z,k}^2}\mathrm{Var}(Z \,|\, Z_k), \qquad \mathrm{Var}(Z \,|\, Z_k) = \Sigma_{11} - \frac{\Sigma_{21}^2}{\Sigma_{22}} = \frac{\Sigma_{11}\sigma_{Z,k}^2}{\sigma_{Z,k}^2 + \Sigma_{11}\mu_{Z,k}^2} = \left(\frac{\delta}{\mathbb{E}(\bar{\beta}^2)} + \rho_{Z,k}\right)^{-1}.$$

Then by (4.6), (4.9) and the Cauchy–Schwarz inequality, we have

$$\rho_{k+1} = \frac{\mu_{k+1}^2}{\sigma_{k+1}^2} = \frac{\mathbb{E}\big(g_k^*(Z_k, Y)\, g_k(Z_k, Y)\big)^2}{\mathbb{E}\big(g_k(Z_k, Y)^2\big)} \le \mathbb{E}\big(g_k^*(Z_k, Y)^2\big),$$

with equality when g_k is a (non-zero) scalar multiple of g_k^*.

Now given μ_k, σ_k for some $k \in \mathbb{N}$, we wish to find $f_k \colon \mathbb{R} \to \mathbb{R}$ such that defining $\Sigma \equiv \Sigma_k$ as in (4.8), the quantity

$$\rho_{Z,k} = \frac{\mu_{Z,k}^2}{\sigma_{Z,k}^2} = \frac{\Sigma_{21}^2 \Sigma_{11}^{-2}}{\Sigma_{22} - \Sigma_{21}^2 \Sigma_{11}^{-1}} = \biggl(\frac{\Sigma_{22}}{\Sigma_{21}^2}\,\Sigma_{11}^2 - \Sigma_{11}\biggr)^{-1}$$

is as large as possible. Since $\Sigma_{11} = \delta^{-1}\mathbb{E}(\bar{\beta}^2)$ is fixed, this amounts to maximising

$$\frac{\Sigma_{21}^2}{\Sigma_{22}} = \frac{\mathbb{E}\bigl(\bar{\beta} f_k(\mu_k\bar{\beta} + \sigma_k\tilde{G}_k)\bigr)^2}{\mathbb{E}\bigl(f_k(\mu_k\bar{\beta} + \sigma_k\tilde{G}_k)^2\bigr)}.$$

Again by the Cauchy–Schwarz inequality (see (3.19) in Section 3.3), this can be done by taking f_k to be any (non-zero) scalar multiple of f_k^* satisfying

$$f_k^*(\mu_k\bar{\beta} + \sigma_k\tilde{G}_k) = \mathbb{E}(\bar{\beta} \,|\, \mu_k\bar{\beta} + \sigma_k\tilde{G}_k), \tag{4.17}$$

in which case $\Sigma_{21} = \Sigma_{22} < \Sigma_{11}$. An exact expression for f_k^* is given by Tweedie's formula (3.17), and if $\bar{\beta} \sim \pi_{\bar{\beta}}$ satisfies the conditions of Lemma 3.8, then f_k^* is Lipschitz. As we saw in (3.18), the choice $f_k = f_k^*$ also minimises the asymptotic mean squared error $\mathbb{E}\bigl\{\bigl(f_k(\mu_k\bar{\beta} + \sigma_k G_k) - \bar{\beta}\bigr)^2\bigr\}$, for given μ_k, σ_k; in other words, f_k^* is the Bayes optimal (i.e. MMSE) denoising function.

By recursively defining $g_k = g_k^*$ (or some scalar multiple thereof) and $f_{k+1} = f_{k+1}^*$ for $k \in \mathbb{N}$ using (4.16) and (4.17), together with corresponding sequences $(\mu_k^*, \sigma_k^*, \mu_{Z,k}^*, \sigma_{Z,k}^* : k \in \mathbb{N})$ of state evolution parameters through (4.6)–(4.7), we obtain a *Bayes-GAMP algorithm* of the form (4.4). A version of this was originally derived by Rangan (2011, Section IV-B) as an approximation to a sum-product loopy belief propagation algorithm. The limiting empirical distributions for the Bayes-GAMP iterates can be obtained from Theorem 4.2, provided that the functions f_k^* and $\tilde{g}_k^* \colon (z,u,v) \mapsto g_k^*(u, h(z,v))$ are all Lipschitz and (G0)–(G5) are satisfied.

Even when $\pi_{\bar{\beta}}$ is not completely known, it can still be possible to tailor the choices of f_k, g_k to wider classes of limiting prior distributions that induce certain types of structure in the signals β. For instance, if we are told that $\beta \in \mathbb{R}^p$ has at most sp non-zero entries for some $s \in (0,1)$ and every $p \equiv p_n$, then as in Section 3.3, we can take each f_k to be a soft-thresholding function $\mathrm{S}_{t_k} \colon u \mapsto \mathrm{sgn}(u)(|u| - t_k)_+$ for some $t_k > 0$. Using an AMP recursion (4.19) of this form (for the linear model in Section 4.3) with appropriately chosen thresholds t_k, Bayati and Montanari (2012) derived exact high-dimensional asymptotics for the Lasso estimator; see Section 4.5.

Spectral initialisation: Under the conditions of Theorem 4.2, it follows from (4.15) and Lemma 4.1 that for each $k \in \mathbb{N}$, the estimates $\hat{\beta}^k$ in the generic GAMP procedure (4.4) satisfy $\langle \hat{\beta}^k, \beta \rangle_p = p^{-1}\sum_{j=1}^p \hat{\beta}_j^k \beta_j \xrightarrow{c} \mathbb{E}\bigl(\bar{\beta} f_k(\mu_k\bar{\beta} + \sigma_k G_k)\bigr) = \mu_{Z,k}\,\mathbb{E}(\bar{\beta}^2)$ as $n, p \to \infty$ with $n/p \to \delta$. To ensure that $\mu_{Z,k} \neq 0$ for some k, and hence that the corresponding $\hat{\beta}^k$ has non-zero asymptotic empirical correlation with the signal β, it is sometimes necessary to start with pilot estimators $\hat{\beta}^0 \in \mathbb{R}^p$ that themselves have the property that $\text{c-lim}_{p\to\infty}\langle \hat{\beta}^0, \beta \rangle_p \neq 0$. Indeed, suppose that the limiting random variables in the state evolution recursion (4.6)–(4.7) are such that

$$\mathbb{E}(\bar{\beta}) = 0 \quad \text{and} \quad \mathbb{E}(Z \,|\, Y) = 0 \quad \text{almost surely}, \tag{4.18}$$

where the latter condition is equivalent to (3.13) in Mondelli and Venkataramanan (2020). Now given estimates $\hat{\beta}^0 \in \mathbb{R}^p$ for which $\text{c-lim}_{n\to\infty}\langle \hat{\beta}^0, \beta \rangle_p = \delta(\Sigma_0)_{21} = 0$, we see from Lemma 4.1 that $\mu_{Z,0} = 0$ and Z_0 is independent of (Z, Y), whence $g_0^*(Z_0, Y) = \mathbb{E}(Z \,|\, Y)/\operatorname{Var}(Z) = 0$ almost surely in (4.16) and $\mu_1 = \mathbb{E}\bigl(g_0^*(Z_0, Y)\, g_0(Z_0, Y)\bigr) = 0$ by (4.9). This means that $\mu_{Z,1}\,\mathbb{E}(\bar{\beta}^2) = \delta(\Sigma_1)_{21} = \mathbb{E}\bigl(\bar{\beta} f_1(\sigma_1 G_1)\bigr) = 0$ by the independence of $\bar{\beta}$ and G_1. Continuing inductively, we conclude that $\mu_k = \mu_{Z,k} = 0$ for all $k \in \mathbb{N}$, irrespective of the choices of g_k, f_{k+1} for $k \in \mathbb{N}_0$, so $\hat{\beta}^k$ is asymptotically uninformative as an estimator of $\beta \in \mathbb{R}^p$ for every $k \in \mathbb{N}_0$.

Thus, while there are some GLMs (such as the linear model in Section 4.3) in which it suffices to take $\hat{\beta}^0 = c\mathbf{1}_p$ for some fixed $c \in \mathbb{R}$, a different initialiser is required when (4.18) holds. We note that the second condition therein is satisfied in the phase retrieval model, where $y_i = h(x_i^\top \beta, \varepsilon_i) = (x_i^\top \beta)^2 + \varepsilon_i$ for $1 \leq i \leq n$, and more generally in all non-identifiable models of the form (4.1) where $h(z, w) = h(-z, w)$ for all z, w (and hence $Q(\cdot \,|\, z) = Q(\cdot \,|\, -z)$ in (4.2) for all z). Indeed, for such functions h, we have $\mathbb{E}(Z \,|\, Y) = \mathbb{E}\big(Z \,|\, h(Z, \bar{\varepsilon})\big) = -\mathbb{E}\big(Z \,|\, h(Z, \bar{\varepsilon})\big)$ and hence $\mathbb{E}(Z \,|\, Y) = 0$ almost surely.

Mondelli and Venkataramanan (2020) established a version of Theorem 4.2 for GAMP algorithms in which $\hat{\beta}^0$ is taken to be a leading eigenvector of $X^\top D X \in \mathbb{R}^{p \times p}$, where $D = \mathrm{diag}\big(g(y_1), \dots, g(y_n)\big) \in \mathbb{R}^{n \times n}$ for some $g \colon \mathbb{R} \to \mathbb{R}$. Since this spectral initialiser is correlated with the random design matrix X, condition (G0) for the original Theorem 4.2 does not hold in general. As mentioned in Section 3.2, the authors overcome this obstacle by analysing a two-phase artificial GAMP iteration in which the first stage effectively approximates $\hat{\beta}^0$ by the power method.

Confidence intervals and p-values: For fixed k and large n, Theorem 4.2 tells us that $\{(\beta_i^k - \mu_k \beta_i)/\sigma_k : 1 \leq i \leq n\}$ behaves approximately like an i.i.d. sample of size n from the $N(0,1)$ distribution. Thus, to carry out inference for β, we can proceed similarly as in Section 3.4, to which we refer the reader for further details. We mention here that if the state evolution parameters μ_k, σ_k are unknown, then they can be estimated consistently by $\hat{\mu}_k := \big(\|\beta^k\|_p^2 - \|\hat{r}^{k-1}\|_n^2\big)^{1/2} / \mathbb{E}(\bar{\beta}^2)^{1/2}$ and $\hat{\sigma}_k := \|\hat{r}^{k-1}\|_n$ provided that $\mathbb{E}(\bar{\beta}^2) > 0$ is known. Indeed, by (4.10) and (4.11) respectively,

$$\begin{aligned} \|\beta^k\|_p^2 &\xrightarrow{c} \mathbb{E}\big((\mu_k \bar{\beta} + \sigma_k G_k)^2\big) = \mathbb{E}(\bar{\beta}^2)\mu_k^2 + \sigma_k^2, \\ \|\hat{r}^{k-1}\|_n^2 = \|g_{k-1}(\theta^{k-1}, y)\|_n^2 &= \|\tilde{g}_{k-1}(\theta, \theta^{k-1}, \varepsilon)\| \xrightarrow{c} \mathbb{E}\big(\tilde{g}_{k-1}(Z, Z_k, \bar{\varepsilon})^2\big) = \mathbb{E}\big(g_{k-1}(Z_k, Y)^2\big) = \sigma_k^2 \end{aligned}$$

for each $k \in \mathbb{N}$ as $n, p \to \infty$ with $n/p \to \delta$.

4.3 AMP for the linear model

Much of the early work on AMP (e.g. Donoho et al., 2009; Bayati and Montanari, 2011, 2012; Krzakala et al., 2012) was centred around the standard linear model

$$y = X\beta + \varepsilon,$$

where $\varepsilon_1, \dots, \varepsilon_n \overset{\mathrm{iid}}{\sim} P_{\bar{\varepsilon}}$ have second moment $\sigma^2 > 0$ and a finite r^{th} moment for some $r \in [2, \infty)$ (or more generally where the empirical distribution $\nu_n(\varepsilon) = n^{-1} \sum_{i=1}^n \delta_{\varepsilon_i}$ converges completely in d_r to $P_{\bar{\varepsilon}}$ as $n \to \infty$). This is a special case of the model (4.1) with $h(z, v) = z + v$.

Given $\hat{r}^{-1} = 0 \in \mathbb{R}^n$, $b_0 \in \mathbb{R}$ and an initial estimator $\hat{\beta}^0 \in \mathbb{R}^p$, the original AMP algorithm of Donoho et al. (2009) and Bayati and Montanari (2011) can be recovered by setting $g_k(u, v) := v - u$ for $u, v \in \mathbb{R}$ in the GAMP recursion (4.4), so that $c_k = \langle g_k'(\theta^k, y) \rangle_n = -1$ and

$$\hat{r}^k = y - X\hat{\beta}^k + b_k \hat{r}^{k-1}, \qquad \hat{\beta}^{k+1} = f_{k+1}(X^\top \hat{r}^k + \hat{\beta}^k), \qquad b_{k+1} = \frac{1}{n} \sum_{j=1}^p f_{k+1}'(\beta_j^{k+1}) \tag{4.19}$$

for $k \in \mathbb{N}_0$. Here, $\hat{r}^k = g_k(\theta^k, y) = y - \theta^k = y - X\hat{\beta}^k + b_k \hat{r}^{k-1}$ is a ‘corrected’ residual at iteration k, and $\beta^{k+1} = X^\top \hat{r}^k + \hat{\beta}^k$ is the effective observation.

State evolution: The GAMP state evolution equations (4.6)–(4.7) simplify to the recursion

$$\mu_k \equiv 1, \qquad \sigma_1^2 = \sigma^2 + \mathbb{E}\big((Z - Z_0)^2\big), \qquad \sigma_{k+1}^2 = \sigma^2 + \frac{1}{\delta} \mathbb{E}\big\{\big(\bar{\beta} - f_k(\bar{\beta} + \sigma_k G_k)\big)^2\big\} \tag{4.20}$$

for $k \in \mathbb{N}$, where $(Z, Z_0) \sim N_2(0, \Sigma_0)$, and $\bar{\beta} \sim \pi_{\bar{\beta}}$ is independent of $G_k \sim N(0,1)$. Note that by (G2), $\sigma_1^2 = \sigma^2 + \text{c-lim}_{n \to \infty} n^{-1} \|\beta - \hat{\beta}^0\|^2$, and that if the pilot estimate of $\beta \in \mathbb{R}^p$ is taken to be $\hat{\beta}^0 = 0 \in \mathbb{R}^p$ for each $p \equiv p_n$, then $Z_0 \equiv 0$ and $\sigma_1^2 = \sigma^2 + \mathbb{E}(Z^2) = \sigma^2 + \delta^{-1} \mathbb{E}(\bar{\beta}^2)$.

Asymptotic estimation error: Under (G0)–(G5) with $r \in [2, \infty)$, the main result of Bayati and Montanari (2011, Theorem 1) on the asymptotic performance of the estimators $\hat{\beta}^k$ in (4.19) can be stated as

$$\sup_{\psi \in \mathrm{PL}_2(r,1)} \left| \frac{1}{p} \sum_{j=1}^{p} \psi(\hat{\beta}_j^k, \beta_j) - \mathbb{E}\bigl\{\psi\bigl(f_k(\bar{\beta} + \sigma_k G_k), \bar{\beta}\bigr)\bigr\} \right| \xrightarrow{c} 0 \quad \text{as } n, p \to \infty \text{ with } n/p \to \delta, \tag{4.21}$$

for each $k \in \mathbb{N}$, where σ_k is as in (4.20). This can be obtained as a special case of Theorem 4.2 and (4.15). Alternatively, (4.21) can be established via a direct reduction to an abstract asymmetric AMP recursion of the type in Section 2.2; see Bayati and Montanari (2011, Section 3.3). This involves writing (4.19) in terms of $e^k := \varepsilon - \hat{r}^k$ and $h^{k+1} := \beta^{k+1} - \beta$, which turn out to be the asymptotically Gaussian 'noise' components of $\hat{r}^k$ and β^{k+1} respectively.

Originally, the d_r convergence result (4.21) was derived under a stronger version of (G1) that assumed d_{2r-2} convergence to limiting distributions $\pi_{\bar{\beta}}, P_{\bar{\varepsilon}}$ with finite $(2r-2)^{th}$ moments. In (G1), we relax this to a more natural d_r condition under which the conclusion still holds; see the first part of Remark 6.4. We also mention that under suitable finite-sample analogues of the conditions above, a complementary finite-sample version of (4.21) was established by Rush and Venkataramanan (2018) in the case $r = 2$; see Remark 6.3.

Link to Bayes-GAMP: If the limiting prior distribution $\pi_{\bar{\beta}}$ is known, then to minimise the effective noise variance σ_{k+1}^2, we can take f_k in (4.19) to be the Bayes optimal f_k^* from (4.17). In general, $g_k \colon (u, v) \mapsto v - u$ does not coincide with g_k^* in (4.16). However, when $P_{\bar{\varepsilon}} = N(0, \sigma^2)$ with $\sigma^2 > 0$, $\hat{\beta}^0 \equiv \hat{\beta}^0(n) = 0$ for every n and $f_k = f_k^*$ for each $k \in \mathbb{N}$, it turns out that (4.19) is an instance of a Bayes-GAMP procedure (with $g_k \propto g_k^*$) that maximises the effective signal-to-noise ratios $\rho_k = (\mu_k/\sigma_k)^2$ and $\rho_{Z,k} = (\mu_{Z,k}/\sigma_{Z,k})^2$ at each iteration. Indeed, in this special case, it can be verified by direct computation that

$$g_k^*(u, v) = c_k\left(\frac{\Sigma_{21}}{\Sigma_{22}} u - v\right) = c_k(u - v) = -c_k g_k(u, v)$$

for each $k \in \mathbb{N}_0$, where $\Sigma \equiv \Sigma_k \in \mathbb{R}^{2\times 2}$ is as in (4.7), with $\Sigma_{21} = \Sigma_{22} = 0 < \Sigma_{11}$ when $k = 0$ and $\Sigma_{21} = \delta^{-1}\mathbb{E}\bigl(\bar{\beta} f_k^*(\mu_k\bar{\beta} + \sigma_k G_k)\bigr) = \delta^{-1}\mathbb{E}\bigl(f_k^*(\mu_k\bar{\beta} + \sigma_k G_k)^2\bigr) = \Sigma_{22} < \Sigma_{11} = \delta^{-1}\mathbb{E}(\bar{\beta}^2)$ by (4.17) when $k \in \mathbb{N}$, and

$$c_k = -\frac{\Sigma_{11} - \Sigma_{22}}{\Sigma_{11} - \Sigma_{22} + \sigma^2} < 0$$

is deterministic. Here, $\delta(\Sigma_{11} - \Sigma_{22}) = \mathbb{E}\bigl\{\bigl(\bar{\beta} - \mathbb{E}(\bar{\beta} \,|\, \mu_k\bar{\beta} + \sigma_k G_k)\bigr)^2\bigr\}$ is the minimum mean squared error for the problem of estimating $\bar{\beta}$ based on $\mu_k\bar{\beta} + \sigma_k G_k$.

4.4 GAMP algorithms for convex optimisation

Given $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ with rows $x_1, \ldots, x_n$, many statistical estimators of β in (4.1) are defined as minimisers of objective functions of the form $\tilde{\beta} \mapsto \mathcal{C}(\tilde{\beta}; X, y) := \sum_{i=1}^n \ell(x_i^\top \tilde{\beta}, y_i) + \sum_{j=1}^p J(\tilde{\beta}_j)$, or equivalently as solutions to constrained optimisation problems of the form

$$\text{minimise} \quad \sum_{i=1}^{n} \ell(\tilde{\theta}_i, y_i) + \sum_{j=1}^{p} J(\tilde{\beta}_j) \quad \text{over } (\tilde{\beta}, \tilde{\theta}) \in \mathbb{R}^p \times \mathbb{R}^n \text{ with } \tilde{\theta} = X\tilde{\beta}, \tag{4.22}$$

where $\ell \colon \mathbb{R}^2 \to \mathbb{R}$ is a loss function and $J \colon \mathbb{R} \to \mathbb{R}$ is a penalty function. In particular, consider a GLM of the form (4.2) in which $y_i \,|\, (x_i, \beta) \sim q(\cdot \,|\, x_i^\top \beta)$ for $1 \le i \le n$, where $q(\cdot \,|\, z)$ is a Lebesgue density on $\mathbb{R}$ for each $z \in \mathbb{R}$. Then the maximum likelihood estimators (MLEs) of β and $\theta = X\beta$ are given by

$$(\hat{\beta}^{\mathrm{MLE}}, \hat{\theta}^{\mathrm{MLE}}) := \operatorname*{argmin}_{\substack{(\tilde{\beta}, \tilde{\theta}) \in \mathbb{R}^p \times \mathbb{R}^n \\ \tilde{\theta} = X\tilde{\beta}}} \sum_{i=1}^{n} -\log q(y_i \,|\, \tilde{\theta}_i).$$

If in addition $\beta_1, \dots, \beta_p \overset{\text{iid}}{\sim} p_{\bar{\beta}}$ for some prior density $p_{\bar{\beta}}$, then the maximum a posteriori (MAP) estimates of β and θ are

$$(\hat{\beta}^{\mathrm{MAP}}, \hat{\theta}^{\mathrm{MAP}}) := \mathop{\mathrm{argmin}}_{\substack{(\tilde{\beta},\tilde{\theta})\in\mathbb{R}^p\times\mathbb{R}^n \\ \tilde{\theta}=X\tilde{\beta}}} \left(\sum_{i=1}^n -\log q(y_i \,|\, \tilde{\theta}_i) + \sum_{j=1}^p -\log p_{\bar{\beta}}(\tilde{\beta}_j)\right).$$

Assuming henceforth that ℓ and J are convex in their first arguments, we will now design a GAMP iteration (4.29) whose fixed points are solutions to the associated optimisation problem (4.22); see Proposition 4.4 below. By exploiting this connection and applying the GAMP theory from Section 4.1, we will explain later how to obtain a statistical payoff in the form of exact high-dimensional asymptotics for estimators defined by (4.22).

To begin the construction, fix two sequences of deterministic scalars $\bar{b}_k > 0$ and $\bar{c}_k < 0$ for $k \in \mathbb{N}_0$. These will later be assigned appropriate values in (4.29) below, but for the time being, we will treat them as generic constants. For $k \in \mathbb{N}_0$, define $\bar{g}_k, g_k \colon \mathbb{R}^2 \to \mathbb{R}$ and $f_{k+1} \colon \mathbb{R} \to \mathbb{R}$ by

$$\bar{g}_k(u,v) := \mathop{\mathrm{argmin}}_{z\in\mathbb{R}} \left\{\ell(z,v) + \frac{1}{2\bar{b}_k}(z-u)^2\right\}, \qquad g_k(u,v) := \frac{\bar{g}_k(u,v) - u}{\bar{b}_k}, \tag{4.23}$$

$$f_{k+1}(w) := \mathop{\mathrm{argmin}}_{z\in\mathbb{R}} \left\{J(z) - \frac{\bar{c}_k}{2}\left(z + \frac{w}{\bar{c}_k}\right)^2\right\}. \tag{4.24}$$

Note that since ℓ and J are assumed to be convex in their first arguments, $\bar{g}_k(u,v)$ and $f_{k+1}(w)$ are well-defined as unique minima of strongly convex functions. The pertinence of this specific choice of g_k, f_{k+1} will become apparent through Proposition 4.4 below and its proof. At this point, it is helpful to recall that for a convex function $\mathrm{M} \colon \mathbb{R} \to \mathbb{R}$ and $\eta > 0$, the associated *proximal operator* $\mathrm{prox}_{\eta\mathrm{M}} \colon \mathbb{R} \to \mathbb{R}$ is given by

$$\mathrm{prox}_{\eta\mathrm{M}}(z) := \mathop{\mathrm{argmin}}_{t\in\mathbb{R}} \left\{\eta\mathrm{M}(t) + \frac{1}{2}(t-z)^2\right\}, \tag{4.25}$$

and moreover that $\mathrm{prox}_{\eta\mathrm{M}}$ is always non-decreasing and 1-Lipschitz (cf. Parikh and Boyd, 2013, Sections 2.3 and 3.1). We see that $\bar{g}_k(u,v) = \mathrm{prox}_{\bar{b}_k\ell(\cdot,v)}(u)$ and $f_{k+1}(w) = \mathrm{prox}_{-J/\bar{c}_k}(-w/\bar{c}_k)$ for $u, v, w \in \mathbb{R}$, so $\bar{g}_k, g_k, f_{k+1}$ are all Lipschitz with constants 1, $\bar{b}_k^{-1}$ and $|\bar{c}_k|^{-1}$ respectively, and hence weakly differentiable with respect to their first arguments. Writing $\bar{g}_k', g_k', f_{k+1}'$ for the corresponding weak derivatives, we have

$$f_{k+1}'(w) \geq 0, \qquad \bar{g}_k'(u,v) \leq 1 \quad \text{and hence} \quad g_k'(u,v) \leq 0 \tag{4.26}$$

for all u, v, w. If in addition ℓ and J are twice continuously differentiable, then $J'(f_{k+1}(w)) - \big(\bar{c}_k f_{k+1}(w) + w\big) = 0$ for each w, so it follows from the implicit function theorem that

$$f_{k+1}'(w) = \big(J''(f_{k+1}(w)) - \bar{c}_k\big)^{-1} \quad \text{and similarly} \quad \bar{g}_k'(u,v) = \big(\bar{b}_k \ell''(f_{k+1}(w)) + 1\big)^{-1} \tag{4.27}$$

for all u, v, w, where ℓ'' denotes the second partial derivative of ℓ with respect to its first argument.

We will now define a GAMP recursion of the form (4.4) as a precursor to the iteration (4.29) that will subsequently be used to analyse the statistical properties of the solutions to the optimisation problem (4.22). Given $\hat{s}^{-1} := 0 \in \mathbb{R}^n$, a fixed $b_0 > 0$ and an initialiser $\hat{\beta}^0 \in \mathbb{R}^p$, inductively define

$$\begin{aligned}
&\theta^k := X\hat{\beta}^k - b_k\hat{s}^{k-1}, && \hat{\theta}^k := \bar{g}_k(\theta^k, y), && c_k := n^{-1}\textstyle\sum_{i=1}^n g_k'(\theta_i^k, y_i), && \hat{s}^k := g_k(\theta^k, y), \\
&\beta^{k+1} := X^\top\hat{s}^k - c_k\hat{\beta}^k, && \hat{\beta}^{k+1} := f_{k+1}(\beta^{k+1}), && b_{k+1} := n^{-1}\textstyle\sum_{j=1}^p f_{k+1}'(\beta_j^{k+1})
\end{aligned} \tag{4.28}$$

for $k \in \mathbb{N}_0$. Note that $\hat{s}^k = (\hat{\theta}^k - \theta^k)/\bar{b}_k$, and that if ℓ and J are convex and twice continuously differentiable with respect to their first arguments, then (4.27) yields

$$c_k = \frac{1}{\bar{b}_k}\left(\frac{1}{n}\sum_{i=1}^n \bar{g}_k'(\theta_i^k, y_i) - 1\right) = -\frac{1}{n}\sum_{i=1}^n \frac{\ell''(\hat{\theta}_i^k, y_i)}{\bar{b}_k\ell''(\hat{\theta}_i^k, y_i) + 1}, \qquad b_{k+1} = \frac{1}{n}\sum_{j=1}^p \frac{1}{J''(\hat{\beta}_j^{k+1}) - \bar{c}_k}.$$

If the hypotheses of Theorem 4.2 are satisfied by a sequence of recursions (4.28), then the limiting empirical distributions of the iterates therein are characterised by the associated state evolution parameters $(\mu_k, \sigma_k, \Sigma_k : k \in \mathbb{N})$ defined through (4.6)–(4.7). Moreover, with $(Z, Z_k) \sim N_2(0, \Sigma_k)$ and $Y = h(Z, \bar{\varepsilon})$ as in Lemma 4.1 for each fixed k, recall from (4.12) that $b_k \xrightarrow{c} \delta^{-1}\,\mathbb{E}\big(f_k'(\mu_k\bar{\beta} + \sigma_k G_k)\big)$ and $c_k \xrightarrow{c} \mathbb{E}\big(g_k'(Z_k, Y)\big)$ as $n, p \to \infty$ with $n/p \to \delta \in (0, \infty)$.

Based on this observation, we will define $\bar{b}_k$ and $\bar{c}_k$ above to coincide with these limiting values, and substitute these deterministic quantities for the random b_k, c_k in (4.28) to obtain the following modified recursion. As before, we start with $\hat{s}^{-1} := 0 \in \mathbb{R}^n$, $\bar{b}_0 > 0$, $\hat{\beta}^0 \in \mathbb{R}^p$, as well as a positive definite $\Sigma_0 \in \mathbb{R}^{2\times 2}$ as in (G2). Given $\hat{\beta}^k, \hat{s}^{k-1}$ and $\bar{b}_k, \Sigma_k$ for a general $k \in \mathbb{N}_0$, we inductively define $\bar{g}_k, g_k$ as in (4.23), along with

$$\begin{aligned} \theta^k &:= X\hat{\beta}^k - \bar{b}_k\hat{s}^{k-1}, & \hat{\theta}^k &:= \bar{g}_k(\theta^k, y), & \bar{c}_k &:= \mathbb{E}\big(g_k'(Z_k, Y)\big), & \hat{s}^k &:= g_k(\theta^k, y), \\ \beta^{k+1} &:= X^\top\hat{s}^k - \bar{c}_k\hat{\beta}^k, & \hat{\beta}^{k+1} &:= f_{k+1}(\beta^{k+1}), & \bar{b}_{k+1} &:= \delta^{-1}\,\mathbb{E}\big(f_{k+1}'(\mu_{k+1}\bar{\beta} + \sigma_{k+1}G_{k+1})\big). \end{aligned} \tag{4.29}$$

In (4.29), we take $(Z, Z_k) \sim N_2(0, \Sigma_k)$ and $Y = h(Z, \bar{\varepsilon})$ as above, and define the state evolution parameters μ_{k+1}, σ_{k+1} as in (4.6) based on g_k, while using $\bar{c}_k$ and (4.24) to specify f_{k+1}. Finally, define Σ_{k+1} in terms of $f_{k+1}, \mu_{k+1}, \sigma_{k+1}$ according to (4.7). We emphasise that the functions $\bar{g}_k, f_{k+1}$ are indeed well-defined through (4.23)–(4.24) for all k since $\bar{b}_k > 0 > \bar{c}_k$ by (4.26) and the fact that $\mathrm{prox}_{\mathrm{M}}$ is non-constant for any convex $\mathrm{M}\colon \mathbb{R} \to \mathbb{R}$.

The iteration (4.29) has two important features that make it a useful theoretical tool. First, Remark 4.3 ensures that its iterates are characterised by the state evolution parameters $(\mu_k, \sigma_k, \Sigma_k : k \in \mathbb{N})$ under the hypotheses of Theorem 4.2. In addition, the following result highlights the significance of (4.29) as an optimisation procedure for the original constrained problem (4.22).

Proposition 4.4 (Rangan et al., 2016, Theorem 1)**.** *In (4.22), suppose that ℓ and J are convex in their first arguments, and define the associated Lagrangian by*

$$L(\tilde{\beta}, \tilde{\theta}, s) := \sum_{i=1}^{n} \ell(\tilde{\theta}_i, y_i) + \sum_{j=1}^{p} J(\tilde{\beta}_j) + s^\top(\tilde{\theta} - X\tilde{\beta}) \tag{4.30}$$

for $\tilde{\beta} \in \mathbb{R}^p$ and $\tilde{\theta}, s \in \mathbb{R}^n$. Then the iterates in (4.29) satisfy

$$\hat{\beta}^{k+1} = \operatorname*{argmin}_{\tilde{\beta}\in\mathbb{R}^p} \Big\{ L(\tilde{\beta}, \hat{\theta}^k, \hat{s}^k) - \frac{\bar{c}_k}{2}\|\tilde{\beta} - \hat{\beta}^k\|^2 \Big\}, \tag{4.31}$$

$$\hat{\theta}^{k+1} = \operatorname*{argmin}_{\tilde{\theta}\in\mathbb{R}^n} \Big\{ L(\hat{\beta}^{k+1}, \tilde{\theta}, \hat{s}^k) + \frac{1}{2\bar{b}_{k+1}}\|\tilde{\theta} - X\hat{\beta}^{k+1}\|^2 \Big\}, \tag{4.32}$$

$$\hat{s}^{k+1} = \hat{s}^k + \frac{(\hat{\theta}^{k+1} - X\hat{\beta}^{k+1})}{\bar{b}_{k+1}}. \tag{4.33}$$

for $k \in \mathbb{N}_0$. Moreover, if $(\beta^*, \theta^*, \hat{\beta}^*, \hat{\theta}^*, \hat{s}^*)$ is a fixed point of (4.29), then $(\hat{\beta}^*, \hat{\theta}^*)$ is a solution to the optimisation problem (4.22), i.e. $\hat{\beta}^* \in \mathrm{argmin}_{\tilde{\beta}\in\mathbb{R}^p}\,\mathcal{C}(\tilde{\beta}; X, y)$.

In fact, the proof we give in Section 6.9 reveals that Proposition 4.4 holds for any choice of deterministic scalars $\bar{b}_k > 0$ and $\bar{c}_k < 0$ in the first column of (4.29), provided that these are also used to define $\bar{g}_k, f_{k+1}$. The characterisation in (4.31)–(4.33) shows that the GAMP algorithm (4.29) is closely related to (but not completely identical to) a 'linearised' Alternating Direction Method of Multipliers (ADMM) procedure (Parikh and Boyd, 2013, Section 4.4.2) for optimising (4.22). Alternating algorithms of this type are particularly well-suited to handling objective functions of the form (4.30) since each minimisation step involves only one of J and ℓ (while (4.33) is a dual update step). The forms of the quadratic penalties in (4.31)–(4.32) ensure that the 'augmented Lagrangians' therein are separable, and hence can be minimised separately in each coordinate of $\tilde{\beta}$ or $\tilde{\theta}$. This is why $\hat{\beta}^{k+1}, \hat{\theta}^{k+1}$ are obtained from $\beta^{k+1}, \theta^{k+1}$ by componentwise applications of $f_{k+1}, \bar{g}_{k+1}$ respectively, whose expressions in (4.23)–(4.24) emerge naturally from (4.31)–(4.32). See also Boyd et al. (2011) for an accessible introduction to ADMM, and Rangan et al. (2016) for further details on the connection between GAMP and conventional convex optimisation algorithms.

Based on Proposition 4.4 and the reasoning above, we might expect the high-dimensional limiting behaviour of the estimators $\hat{\beta}^* \in \operatorname{argmin}_{\tilde{\beta}\in\mathbb{R}^p} \mathcal{C}(\tilde{\beta}; X, y)$ to be governed by some fixed point of the state evolution for (4.29) (if it exists). To prove this, we might hope to be able to establish convergence of both the GAMP iteration (4.29) and its state evolution to their respective fixed points (in the sense of (4.36) below). We conclude this subsection by setting out a general strategy along these lines. In Sections 4.5–4.7, we will go on to demonstrate that it unifies existing derivations of high-dimensional asymptotic results for the Lasso, and M-estimators in the linear model and logistic regression model.

Step 1: For given ℓ and J (and fixed n and $p \equiv p_n$), find a fixed point of (4.29) *together with its state evolution*, satisfying

$$\begin{aligned} &\theta^* := X\hat{\beta}^* - \bar{b}_*\hat{s}^*, && \hat{\theta}^* := \bar{g}_*(\theta^*, y), && \bar{c}_* := \mathbb{E}\big(g_*'(Z_*, Y)\big), && \hat{s}^* := g_*(\theta^*, y), \\ &\beta^* := X^\top\hat{s}^* - \bar{c}_*\hat{\beta}^*, && \hat{\beta}^* := f_*(\beta^*), && \bar{b}_* := \delta^{-1}\,\mathbb{E}\big(f_*'(\mu_*\bar{\beta} + \sigma_* G_*)\big). && \end{aligned} \tag{4.34}$$

Here, $f_*, \bar{g}_*, g_*$ are defined in terms of $\bar{b}_* > 0$, $\bar{c}_* < 0$ as in (4.23)–(4.24), with $(Z, Z_*) \sim N_2(0, \Sigma_*)$ and $Y = h(Z, \bar{\varepsilon})$, while $G_* \sim N(0,1)$ is independent of $\bar{\beta} \sim \pi_{\bar{\beta}}$ and $\mu_*, \sigma_*, \Sigma_*, f_*, g_*$ satisfy (4.6)–(4.7). In each of the subsequent examples, the system (4.34) reduces to a smaller set of (non-linear) equations. The existence and uniqueness of a state evolution fixed point usually needs to be verified on a case-by-case basis, and may depend on the values of parameters such as the limiting sampling ratio δ and the asymptotic signal strength $\mathbb{E}(\beta^2)/\delta$ (the variance of Z above).

Step 2: If Step 1 yields suitable $f_*, \bar{g}_*, g_*, \bar{b}_*, \bar{c}_*$, then consider the following 'stationary' version of (4.29) for each n and $p \equiv p_n$:

$$\begin{aligned} \theta^k &:= X\hat{\beta}^k - \bar{b}_*\hat{s}^{k-1}, & \hat{\theta}^k &:= \bar{g}_*(\theta^k, y), & \hat{s}^k &:= g_*(\theta^k, y), \\ \beta^{k+1} &:= X^\top\hat{s}^k - \bar{c}_*\hat{\beta}^k, & \hat{\beta}^{k+1} &:= f_*(\beta^{k+1}). & & \end{aligned} \tag{4.35}$$

Henceforth, we will use (4.35) as a theoretical device rather than as a practical algorithm, which gives us the flexibility to initialise it with $\hat{s}^{-1} = 0 \in \mathbb{R}^n$ and an 'oracle' $\hat{\beta}^0 = f_*(\mu_*\beta + \sigma_*\xi) \in \mathbb{R}^p$, where $\xi \sim N_p(0, I_p)$ is independent of the signal $\beta \in \mathbb{R}^p$. This is a convenient choice because it ensures that $\Sigma_0 = \Sigma_*$ and hence that the state evolution for (4.35) is stationary, i.e. $\mu_k = \mu_*$, $\sigma_k = \sigma_*$ and $\Sigma_k = \Sigma_*$ for all $k \in \mathbb{N}$. In addition, as $n, p \to \infty$ with $n/p \to \delta$ under (G1), the d_2 limit of the empirical distribution of the entries of $\hat{\beta}^0$ is the distribution of $f_*(\mu_*\bar{\beta} + \sigma_* G_*)$ by construction, and under the hypotheses of Theorem 4.2, this is also true of $\hat{\beta}^k$ for each *fixed* $k \in \mathbb{N}$ by Remark 4.3. The remaining technical challenge to establish the same distributional limit for the fixed point $\hat{\beta}^*$, which solves the optimisation problem (4.22) by Proposition 4.4.

Step 3: Show that the estimates $\hat{\beta}^k$ in (4.35) converge to $\hat{\beta}^* \in \operatorname{argmin}_{\tilde{\beta}\in\mathbb{R}^p} \mathcal{C}(\tilde{\beta}; X, y)$ in the sense that

$$\lim_{k\to\infty} \operatorname*{c-lim}_{p\to\infty} \frac{\|\hat{\beta}^k - \hat{\beta}^*\|^2}{p} = 0. \tag{4.36}$$

In the examples in Sections 4.5–4.7, this is achieved by first establishing a 'Cauchy property'

$$\operatorname*{c-lim}_{p\to\infty} \frac{\|\hat{\beta}^{k+1} - \hat{\beta}^k\|^2}{p} = 0, \qquad \operatorname*{c-lim}_{n\to\infty} \frac{\left\|\hat{s}^{k+1} - \hat{s}^k\right\|^2}{n} = 0$$

for each k (using the limiting covariance structure mentioned after Theorem 4.2), and then proving that for large k and p, the original convex cost function $\tilde{\beta} \mapsto \mathcal{C}(\tilde{\beta}; X, y)$ is approximately minimised by $\hat{\beta}^k$ in the following sense: if $\hat{\gamma}^k \in \mathbb{R}^p$ belongs to the subgradient of $\mathcal{C}(\cdot\,; X, y)$ at $\hat{\beta}^k$ for $k \in \mathbb{N}$ and $p \equiv p_n$, then

$$\lim_{k\to\infty} \operatorname*{c-lim}_{p\to\infty} \frac{\|\hat{\gamma}^k\|^2}{p} = 0. \tag{4.37}$$

If $\mathcal{C}(\cdot\,; X, y)$ is strongly convex (on a subset of its domain that contains $\hat\beta^k, \hat\beta^*$) with high probability, then the desired conclusion (4.36) follows readily from (4.37) and the basic inequality $\mathcal{C}(\hat\beta^*; X, y) \leq \mathcal{C}(\hat\beta^k; X, y)$; see (98)–(100) in Donoho and Montanari (2016). Otherwise (as in the case of the Lasso in Section 4.5), further work must be done to show that in a random design setting, it is vanishingly unlikely that $\|\hat\gamma^k\|_p$ is small but $\|\hat\beta^k - \hat\beta^*\|_p$ is large (cf. Bayati and Montanari, 2012, Theorem 1.8 and Lemma 3.1).

4.5 AMP for the Lasso

In high-dimensional linear models $y = X\beta + \varepsilon$, the Lasso (Tibshirani, 1996) is a popular method for obtaining sparse estimates of $\beta \in \mathbb{R}^p$ via ℓ_1-penalised least squares. Given $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$ and a regularisation parameter $\lambda > 0$, the Lasso estimator is defined by

$$\hat\beta^{\mathrm{L},\lambda} \in \operatorname*{argmin}_{\tilde\beta\in\mathbb{R}^p} \left\{\frac{1}{2}\|y - X\tilde\beta\|^2 + \lambda\|\tilde\beta\|_1\right\}. \tag{4.38}$$

In the random design setting of (G0) and (G1), Bayati and Montanari (2012) derived an exact expression (4.46) for the asymptotic estimation error of $\hat\beta^{\mathrm{L},\lambda}$ as $n, p \to \infty$ with $n/p \to \delta \in (0, \infty)$. By following the GAMP recipe in Section 4.4, we will show how to design and calibrate an AMP iteration that is central to the proof of their main result (Theorem 4.5 below).

To begin with, note that $\hat\beta^{\mathrm{L},\lambda}$ solves a convex optimisation problem of the form (4.22) with $\ell\colon (u, v) \mapsto (u - v)^2/2$ and $J\colon x \mapsto \lambda|x|$. For $k \in \mathbb{N}_0$, the corresponding $\bar g_k, g_k, f_{k+1}$ in (4.23)–(4.24) are given by

$$\bar g_k(u, v) = \frac{u + \bar b_k v}{1 + \bar b_k}, \qquad g_k(u, v) = \frac{v - u}{1 + \bar b_k}, \qquad f_{k+1}(w) = -\,\mathrm{ST}_{\lambda/\bar c_k}\left(-\frac{w}{\bar c_k}\right) = -\frac{\mathrm{ST}_\lambda(w)}{\bar c_k}, \tag{4.39}$$

where as in Section 3.3, we denote by ST_t the soft-thresholding function $w \mapsto \mathrm{sgn}(w)(|w| - t)_+$ for $t > 0$. Given $\hat r^{-1} = 0 \in \mathbb{R}^n$, $\tilde b_0 \equiv \bar b_0 > 0$ and $\hat\beta^0 \in \mathbb{R}^p$, the resulting GAMP algorithm (4.29) can be succinctly written as

$$\hat r^k = y - X\hat\beta^k + \tilde b_k \hat r^{k-1}, \qquad \hat\beta^{k+1} = \mathrm{ST}_{t_{k+1}}\big(X^\top \hat r^k + \hat\beta^k\big) \qquad \text{for } k \in \mathbb{N}_0, \tag{4.40}$$

where $\hat r^k := y - \hat\theta^k$. Observe that (4.40) is (asymptotically equivalent to) an instance of the AMP recursion (4.19) in Section 4.3 for the linear model, whose state evolution formula is given by (4.20), with $\mu_k = 1$ for all k. By (4.29) and (4.39), the deterministic scalars $\tilde b_k := \bar b_k/(1 + \bar b_{k-1}) > 0$ and $t_{k+1} := \lambda(1 + \bar b_k) = -\lambda/\bar c_k > 0$ in (4.40) are related to each other and the state evolution parameters σ_k^2 via

$$\begin{aligned}
\sigma_1^2 &= \sigma^2 + \mathbb{E}\big((Z - Z_0)^2\big), \quad t_1 = \lambda(1 + \tilde b_0), & \tilde b_k &= \frac{\mathbb{E}\big(\mathrm{ST}'_{t_k}(\bar\beta + \sigma_k G_k)\big)}{\delta} = \frac{\mathbb{P}\big(|\bar\beta + \sigma_k G_k| > t_k\big)}{\delta}, \\
\sigma_{k+1}^2 &= \sigma^2 + \frac{\mathbb{E}\big\{\big(\bar\beta - \mathrm{ST}_{t_k}(\bar\beta + \sigma_k G_k)\big)^2\big\}}{\delta}, & t_{k+1} &= \lambda + \tilde b_k t_k = \lambda + \frac{t_k\,\mathbb{P}\big(|\bar\beta + \sigma_k G_k| > t_k\big)}{\delta}
\end{aligned} \tag{4.41}$$

for $k \in \mathbb{N}$. Here, $\bar\beta \sim \pi_{\bar\beta}$ and $G_k \sim N(0, 1)$ are independent, $(Z, Z_0) \sim N_2(0, \Sigma_0)$, and $\sigma^2 > 0$ is the second moment of $P_{\bar\varepsilon}$.

Proceeding as in **Step 1** in Section 4.4, we now seek a fixed point $(\hat r^*, \hat\beta^*, \tilde b_*, \sigma_*, t_* > 0)$ of (4.40)–(4.41) satisfying

$$\hat r^* = y - X\hat\beta^* + \tilde b_* \hat r^*, \qquad \tilde b_* = \frac{t_* - \lambda}{t_*}, \qquad \hat\beta^* = \mathrm{ST}_{t_*}\big(X^\top \hat r^* + \hat\beta^*\big), \tag{4.42}$$

$$\sigma_*^2 = \sigma^2 + \frac{\mathbb{E}\big\{\big(\bar\beta - \mathrm{ST}_{t_*}(\bar\beta + \sigma_* G_*)\big)^2\big\}}{\delta}, \qquad t_* = \lambda\left(1 - \frac{\mathbb{P}(|\bar\beta + \sigma_* G_*| > t_*)}{\delta}\right)^{-1}, \tag{4.43}$$

where $\bar{\beta} \sim \pi_{\bar{\beta}}$ and $G_* \sim N(0,1)$ are independent. Noting that the condition (4.42) simplifies to $\hat{\beta}^* = \mathrm{ST}_{t_*}\bigl(\hat{\beta}^* + t_*\lambda^{-1}X^\top(y - X\hat{\beta}^*)\bigr)$, we can either apply Proposition 4.4 or verify the Karush–Kuhn–Tucker (KKT) conditions directly to deduce that $\hat{\beta}^*$ is a Lasso solution satisfying (4.38).

The next task is to show that for any $\lambda > 0$ in (4.38) and $\delta, \sigma > 0$, there exist unique solutions $\sigma_* \equiv \sigma_*(\lambda,\delta,\sigma) > 0$ and $t_* \equiv t_*(\lambda,\delta,\sigma)$ to the non-linear equations in (4.43). To this end, Bayati and Montanari (2012, Proposition 1.3) first verified that for fixed $\alpha > 0$, there is a unique $\tilde{\sigma}_\alpha \equiv \tilde{\sigma}_\alpha(\delta,\sigma)$ satisfying

$$\tilde{\sigma}_\alpha^2 = \sigma^2 + \frac{\mathbb{E}\bigl\{\bigl(\bar{\beta} - \mathrm{ST}_{\alpha\tilde{\sigma}_\alpha}(\bar{\beta} + \tilde{\sigma}_\alpha G_*)\bigr)^2\bigr\}}{\delta}$$

provided that

$$\upsilon(\alpha) := (1+\alpha^2)\Phi(-\alpha) - \alpha\phi(\alpha) < \frac{\delta}{2}, \tag{4.44}$$

where ϕ and Φ denote the standard Gaussian density and distribution functions respectively. Since $\upsilon\colon \mathbb{R} \to \mathbb{R}$ is a strictly decreasing continuous function with range $(0,\infty)$, (4.44) holds for all positive $\alpha > \upsilon^{-1}(\delta/2)$. In addition, some elementary calculus shows that for some $\alpha_0 \equiv \alpha_0(\delta,\sigma) \geq \upsilon^{-1}(\delta/2)$, the map

$$\alpha \mapsto \Lambda_{\delta,\sigma}(\alpha) := \alpha\tilde{\sigma}_\alpha\left(1 - \frac{\mathbb{P}(|\bar{\beta} + \tilde{\sigma}_\alpha G_*| > \alpha\tilde{\sigma}_\alpha)}{\delta}\right)$$

is a continuous bijection from (α_0,∞) to $(0,\infty)$ (Bayati and Montanari, 2012, Proposition 1.4 and Corollary 1.7), so that for any $\lambda > 0$, there is a unique $\alpha_* \equiv \alpha_*(\lambda,\delta,\sigma) > \alpha_0$ such that $\lambda = \Lambda_{\delta,\sigma}(\alpha_*)$. It follows from this that $\sigma_* = \tilde{\sigma}_{\alpha_*}$ and $t_* = \alpha_*\sigma_*$ are the unique solutions to (4.43).

For $n \in \mathbb{N}$ and $p \equiv p_n$, the resulting 'stationary' AMP iteration (4.35) in **Step 2** in Section 4.4 takes the form

$$\hat{r}^k = y - X\hat{\beta}^k + \tilde{b}_*\hat{r}^{k-1}, \qquad \hat{\beta}^{k+1} = \mathrm{ST}_{t_*}\bigl(X^\top\hat{r}^k + \hat{\beta}^k\bigr) \qquad \text{for } k \in \mathbb{N}_0, \tag{4.45}$$

where $\hat{r}^{-1} = 0 \in \mathbb{R}^n$, $\tilde{b}_* = \delta^{-1}\,\mathbb{P}(|\bar{\beta} + \sigma_* G_*| > t_*)$, and $\hat{\beta}^0 = \mathrm{ST}_{t_*}(\beta + \sigma_*\xi) \in \mathbb{R}^p$ is an oracle initialiser with $\xi \sim N_p(0, I_p)$ taken to be independent of the signal $\beta \in \mathbb{R}^p$. Under the hypotheses of Theorem 4.2, it follows from Remark 4.3 and (4.15) that for each fixed $k \in \mathbb{N}_0$, the empirical distribution of the entries of $\hat{\beta}^k \equiv \hat{\beta}^k(n)$ converges completely in d_2 to the distribution of $\mathrm{ST}_{t_*}(\bar{\beta} + \sigma_* G_*)$ as $n, p \to \infty$ with $n/p \to \delta$.

Theorem 4.5 below asserts that the same asymptotic conclusion holds for the fixed point $\hat{\beta}^*$ of (4.45), which is a Lasso solution by virtue of (4.42). The additional technical challenge in its proof is to show that the AMP iterates $\hat{\beta}^k$ in (4.45) actually converge to a fixed point in the sense of (4.36), when we take $n, p \to \infty$ followed by $k \to \infty$ (Bayati and Montanari, 2012, Theorem 1.8).[†] This constitutes **Step 3** in Section 4.4, and as mentioned there, the arguments involved turn out to be highly non-trivial in this case because the Lasso objective function in (4.38) is not strongly convex.

Theorem 4.5 (Bayati and Montanari, 2012, Theorem 1.5)**.** *Consider a sequence of linear models $y = X\beta + \varepsilon$ satisfying (G0) and (G1) for $r = 2$ as $n, p \to \infty$ with $n/p \to \delta \in (0,\infty)$. Suppose that the limiting prior distribution $\pi_{\bar{\beta}}$ satisfies $\pi_{\bar{\beta}}(\{0\}) > 0$, so that an asymptotically non-vanishing proportion of the entries of $\beta \in \mathbb{R}^p$ are equal to 0. For $\lambda > 0$, let $\hat{\beta}^{\mathrm{L},\lambda} \in \mathbb{R}^p$ be a Lasso estimator (4.38) for each $p \equiv p_n$, and let $\sigma_* \equiv \sigma_*(\lambda,\delta,\sigma) > 0$ and $t_* \equiv t_*(\lambda,\delta,\sigma) > 0$ be the unique solutions to (4.43). Then*

$$\sup_{\psi \in \mathrm{PL}_2(2,1)} \left|\frac{1}{p}\sum_{j=1}^p \psi(\hat{\beta}_j^{\mathrm{L},\lambda}, \beta_j) - \mathbb{E}\bigl\{\psi\bigl(\mathrm{ST}_{t_*}(\bar{\beta} + \sigma_* G), \bar{\beta}\bigr)\bigr\}\right| \xrightarrow{c} 0 \tag{4.46}$$

[†] Bayati and Montanari (2012) originally established this result for a AMP recursion (4.40) initialised with $\hat{\beta}^0 = 0$, in which the thresholds are defined instead by $t_{k+1} = \alpha_*\sigma_{k+1}$ in (4.41) with $\alpha_* = \alpha_*(\lambda,\delta,\sigma)$ as above, and the state evolution sequence (σ_k) is non-constant but converges to σ_*. Their analysis yields the same conclusion for (4.45), and also shows that (4.48) holds even though $\psi\colon (u,v) \mapsto \mathbb{1}_{\{u \neq 0\}}$ is discontinuous.

as $n, p \to \infty$ with $n/p \to \delta$, where $\bar{\beta} \sim \pi_{\bar{\beta}}$ is independent of $G \sim N(0,1)$. In particular, the asymptotic mean squared error of the Lasso estimator is given by

$$\operatorname*{c-lim}_{p\to\infty} \frac{\|\hat{\beta}^{\mathrm{L},\lambda} - \beta\|^2}{p} = \mathbb{E}\bigl\{\bigl(\bar{\beta} - \mathrm{ST}_{t_*}(\bar{\beta} + \sigma_* G)\bigr)^2\bigr\} = \delta(\sigma_*^2 - \sigma^2). \tag{4.47}$$

We emphasise once again the complex, non-linear dependence of σ_* in (4.46) on the asymptotic sparsity level $\pi_{\bar{\beta}}(\{0\})$ and $\lambda, \delta, \sigma > 0$ through (4.43), and also the fact the asymptotic guarantees of Theorem 4.5 hold for a *fixed* value of the regularisation parameter $\lambda > 0$. Mousavi et al. (2018) showed that the asymptotic mean squared error of $\hat{\beta}^{\mathrm{L},\lambda}$ in (4.47) is a quasi-convex function of λ (i.e. decreasing on $(0, \lambda^*]$ and increasing on $[\lambda^*, \infty)$ for some $\lambda^* > 0$), and moreover that

$$\operatorname*{c-lim}_{p\to\infty} \frac{\|\hat{\beta}^{\mathrm{L},\lambda}\|_0}{p} \equiv \operatorname*{c-lim}_{p\to\infty} \frac{1}{p}\sum_{j=1}^{p} \mathbb{1}_{\{\hat{\beta}_j^{\mathrm{L},\lambda} \neq 0\}} = \mathbb{P}\bigl(\mathrm{ST}_{t_*}(\bar{\beta} + \sigma_* G_*) \neq 0\bigr) = \mathbb{P}(|\bar{\beta} + \sigma_* G_*| > t_*) = \delta\tilde{b}_*(\lambda, \delta, \sigma) \tag{4.48}$$

is a decreasing function of λ, as might be intuitively expected.

When the Lasso is used to perform variable selection (possibly with an adaptive choice of λ), Su et al. (2017) established a tradeoff between the false discovery proportion and false negative proportion along the regularisation path $\lambda \mapsto \hat{\beta}^{\mathrm{L},\lambda}$ in the high-dimensional asymptotic regime above. To this end, by extending the results of Bayati and Montanari (2012), they proved that these two quantities converge *uniformly* to deterministic limits over $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, for any $0 < \lambda_{\min} < \lambda_{\max}$.

Li and Wei (2021) derived precise asymptotics in the over-parametrised regime $p < n$ for the *minimum ℓ_1 norm interpolator*

$$\hat{\beta}^{\mathrm{Int}} := \operatorname*{argmin}_{\tilde{\beta} \in \mathbb{R}^p,\, y = X\tilde{\beta}} \|\tilde{\beta}\|_1,$$

which corresponds to taking the limit $\lambda \searrow 0$ in the Lasso problem. To achieve this, they extended the existing machinery outlined above to sequences of AMP iterations for Lasso estimators with decreasing values of the regularisation parameter $\lambda \equiv \lambda_n \searrow 0$. Their analysis reveals that in the regime $n/p \to \delta < 1$, the asymptotic generalisation error of $\hat{\beta}^{\mathrm{Int}}$ can exhibit several phases of descent and ascent as δ decreases (i.e. as the model complexity increases). This intriguing *multiple descent* behaviour of the generalisation risk curve has been observed empirically for a variety of popular procedures in statistics and machine learning, including random forests and neural networks (e.g. Belkin et al., 2019; Geiger et al., 2019; Advani et al., 2020; Nakkiran et al., 2021). The theoretical study of this phenomenon is a very active area of current research (e.g. Bartlett et al., 2020; Belkin et al., 2020; d'Ascoli et al., 2020; Liang and Rakhlin, 2020; Mei and Montanari, 2020; Hastie et al., 2022); see Dar et al. (2021) for a survey of recent developments.

Remark 4.6. The SLOPE estimator (Bogdan et al., 2015; Su and Candès, 2016; Bellec et al., 2018) is a generalisation of the Lasso that solves a regularised least squares problem in which the penalty is a *sorted* ℓ_1 norm: for $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, define

$$\hat{\beta}^{\mathrm{SLOPE}}(\lambda_1, \dots, \lambda_p) \in \operatorname*{argmin}_{\tilde{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{2}\|y - X\tilde{\beta}\|^2 + \sum_{j=1}^{p} \lambda_j |\tilde{\beta}|_{(j)} \right\}, \tag{4.49}$$

where $|\tilde{\beta}|_{(1)} \geq |\tilde{\beta}|_{(2)} \geq \ldots \geq |\tilde{\beta}|_{(p)}$ are the absolute values of the entries of $\tilde{\beta}$ arranged in decreasing order. This is a convex optimisation problem that produces sparse solutions like the Lasso, but offers more flexibility due to the choices available for $\lambda_1, \dots, \lambda_p$. For example, SLOPE can be used to control the false discovery rate in variable selection via a judicious choice of these regularisation parameters. Note however that when the λ_j are distinct, the optimisation problem (4.49) is not of the form (4.22) since the SLOPE penalty is not an additively separable function of the components of $\tilde{\beta}$. Consequently, the GAMP construction (4.29) in Section 4.4 is not applicable to this setting.

Nevertheless, Bu et al. (2021) show that an appropriately tuned AMP algorithm converges to the SLOPE solution in the sense of (4.36), under assumptions similar to those for Theorem 4.5. This

AMP iteration for SLOPE is somewhat similar to that for the Lasso, the main difference being that the soft-thresholding function in (4.40) is replaced by the proximal operator associated with the SLOPE penalty. This proximal operator is non-separable (i.e. does not act componentwise on its vector input), which is why the analysis is based on master theorems recently obtained by Berthier et al. (2020) for AMP recursions with non-separable denoising functions.

4.6 AMP for M-estimation in the linear model

Consider again the linear model $y = X\beta + \varepsilon$ from Section 4.3, and define an M-estimator of $\beta \in \mathbb{R}^p$ by

$$\hat{\beta}^{\mathrm{M}} \in \operatorname*{argmin}_{\tilde{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} \mathrm{M}(y_i - x_i^\top \tilde{\beta}) \tag{4.50}$$

for some convex $\mathrm{M}\colon \mathbb{R} \to \mathbb{R}$ that is bounded below. The existence of $\hat{\beta}^{\mathrm{M}}$ is guaranteed if for example M is strongly convex. If $\varepsilon_1, \dots, \varepsilon_n \overset{\text{iid}}{\sim} f_{\bar{\varepsilon}}$ for some known (strictly positive log-concave) density $f_{\bar{\varepsilon}}$, then taking $\mathrm{M} = -\log f_{\bar{\varepsilon}}$ in (4.50) yields a maximum likelihood estimator of β; see Dümbgen et al. (2011, Section 3) for a maximum likelihood approach to estimating β when $f_{\bar{\varepsilon}}$ is unknown. Other popular choices of M include squared error loss $w \mapsto w^2$, Huber loss $w \mapsto w^2 \mathbb{1}_{\{|w| \le B\}} + (2|w| - B)B\mathbb{1}_{\{|w|>B\}}$ (for robust regression) with $B > 0$, and quantile loss $w \mapsto \tau w - \mathbb{1}_{\{w<0\}}$ (for quantile regression) with $\tau \in (0,1)$. In a classical setting where the dimension p is fixed, $x_1, \dots, x_n \overset{\text{iid}}{\sim} P_X$ on $\mathbb{R}^p$ and $\varepsilon_1, \dots, \varepsilon_n \overset{\text{iid}}{\sim} P_{\bar{\varepsilon}}$ on $\mathbb{R}$ for all n, Huber (1964, 1973) proved that

$$\sqrt{n}(\hat{\beta}^{\mathrm{M}} - \beta) \xrightarrow{d} N_p(0, \Sigma^{\mathrm{M}}) \quad \text{as } n \to \infty, \quad \text{with} \quad \Sigma^{\mathrm{M}} := \frac{\int_{\mathbb{R}} (\mathrm{M}')^2 \, dP_{\bar{\varepsilon}}}{\bigl(\int_{\mathbb{R}} \mathrm{M}'' \, dP_{\bar{\varepsilon}}\bigr)^2} \left(\int_{\mathbb{R}^p} x x^\top \, dP_X \right)^{-1}, \tag{4.51}$$

under appropriate regularity conditions on M and the *score function* $\mathrm{S} := \mathrm{M}'$; see also Huber and Ronchetti (2009) and van der Vaart (1998, Example 5.28). When $\bar{\varepsilon} \sim P_{\bar{\varepsilon}}$ has a differentiable density $f_{\bar{\varepsilon}}$, it follows from the Cauchy–Schwarz inequality that the variance functional

$$V(\mathrm{S}; \bar{\varepsilon}) := \frac{\mathbb{E}\bigl(\mathrm{S}(\bar{\varepsilon})^2\bigr)}{\mathbb{E}\bigl(\mathrm{S}'(\bar{\varepsilon})\bigr)^2} = \frac{\int_{\mathbb{R}} (\mathrm{M}')^2 \, dP_{\bar{\varepsilon}}}{\bigl(\int_{\mathbb{R}} \mathrm{M}'' \, dP_{\bar{\varepsilon}}\bigr)^2} \tag{4.52}$$

that appears in (4.51) is bounded below by the Fisher information $I(P_{\bar{\varepsilon}}) := \int_{\mathbb{R}} (f_{\bar{\varepsilon}}'/f_{\bar{\varepsilon}})^2 \, dP_{\bar{\varepsilon}}$, with equality when $\mathrm{M} = -\log f_{\bar{\varepsilon}}$ (in which case the maximum likelihood estimator $\hat{\beta}^{\mathrm{M}}$ is asymptotically efficient).

In contrast to (4.51), Donoho and Montanari (2016) showed that the M-estimator $\hat{\beta}^{\mathrm{M}}$ suffers from variance inflation (and cannot be asymptotically efficient) in high-dimensional regimes where $n, p \to \infty$ with $n/p \to \delta \in (1, \infty)$. AMP machinery plays a pivotal role in the analysis that leads to their main result (stated as Theorem 4.7 below), and as in Section 4.5, we will now present the main steps within the context of the GAMP framework of Sections 4.1 and 4.4.

Observing that the convex optimisation problem in (4.50) is an instance of (4.22) with $\ell\colon (u,v) \mapsto \mathrm{M}(u - v)$ and $J \equiv 0$ (i.e. no penalty term), we first write down an associated GAMP algorithm (4.53) based on the general construction in Section 4.4. For $\eta > 0$, define a 'smoothed' version of $\eta\mathrm{M}$ by

$$\mathrm{M}_\eta(z) := \min_{t \in \mathbb{R}} \left\{ \eta \mathrm{M}(t) + \frac{1}{2}(t - z)^2 \right\}$$

for $z \in \mathbb{R}$. (The function $\eta^{-1}\mathrm{M}_\eta$ is called a *Moreau envelope* of M.) We note here that $\operatorname{prox}_{\eta\mathrm{M}}(z)$ in (4.25) is the unique t that achieves this minimum for each $z \in \mathbb{R}$, and also that M_η is convex and differentiable with $\mathrm{S}_\eta(z) := (\mathrm{M}_\eta)'(z) = z - \operatorname{prox}_{\eta\mathrm{M}}(z)$ for all z; see for example Rockafellar (1997, Theorem 31.5) and Parikh and Boyd (2013, Section 3.2). Moreover, S_η is non-decreasing and 1-Lipschitz (cf. Parikh and Boyd, 2013, Sections 2.3 and 3.1).

For $k \in \mathbb{N}_0$, the functions $\bar{g}_k, g_k, f_{k+1}$ in (4.23)–(4.24) are given by

$$\bar{g}_k(u,v) = v - \mathrm{prox}_{\bar{b}_k \mathrm{M}}(v-u) = u + \mathrm{S}_{\bar{b}_k}(v-u), \qquad g_k(u,v) = \frac{\mathrm{S}_{\bar{b}_k}(v-u)}{\bar{b}_k}, \qquad f_{k+1}(w) = -\frac{w}{\bar{c}_k}.$$

Given $\bar{b}_0 > 0$, $\hat{\beta}^0 \in \mathbb{R}^p$ and $\hat{r}^0 := y - X\hat{\beta}^0$, we now write the GAMP algorithm (4.29) in terms of $\hat{r}^k = y - \theta^k$ and $\hat{\beta}^{k+1} = f_{k+1}(\beta^{k+1}) = -\beta^{k+1}/\bar{c}_k$, and obtain the recursion

$$\hat{\beta}^{k+1} = \frac{\delta \bar{b}_{k+1}}{\bar{b}_k} X^\top \mathrm{S}_{\bar{b}_k}(\hat{r}^k) + \hat{\beta}^k, \qquad \hat{r}^{k+1} = y - X\hat{\beta}^{k+1} + \frac{\bar{b}_{k+1}}{\bar{b}_k}\mathrm{S}_{\bar{b}_k}(\hat{r}^k) \qquad \text{for } k \in \mathbb{N}_0, \tag{4.53}$$

where $\bar{b}_{k+1} = -1/(\delta \bar{c}_k)$. Moreover, expressing the state evolution recursion (4.6)–(4.7) for (4.29) in terms of $\tilde{\mu}_k := \delta \bar{b}_k \mu_k$, $\tilde{\sigma}_k := \delta \bar{b}_k \sigma_k$ and $\tau_k := \mathbb{E}\big((Z - Z_k)^2\big)^{1/2}$ with $(Z, Z_k) \sim N(0, \Sigma_k)$, we have $\tau_0 = \mathbb{E}\big((Z - Z_0)^2\big)^{1/2}$ and

$$\begin{aligned} \tilde{\mu}_{k+1} &= \delta\,\mathbb{E}\big(\mathrm{S}'_{\bar{b}_k}(\bar{\varepsilon} + \tau_k G_k)\big), & \tilde{\sigma}^2_{k+1} &= \delta^2\,\mathbb{E}\big(\mathrm{S}_{\bar{b}_k}(\bar{\varepsilon} + \tau_k G_k)^2\big), \\ \bar{b}_{k+1} &= -\frac{1}{\delta \bar{c}_k} = \frac{\bar{b}_k}{\delta\,\mathbb{E}\big(\mathrm{S}'_{\bar{b}_k}(\bar{\varepsilon} + \tau_k G_k)\big)}, & \tau^2_{k+1} &= \frac{\mathbb{E}(\bar{\beta}^2)(\tilde{\mu}_{k+1} - 1)^2 + \tilde{\sigma}^2_{k+1}}{\delta} \end{aligned} \tag{4.54}$$

for $k \in \mathbb{N}_0$, where $\bar{\varepsilon} \sim P_{\bar{\varepsilon}}$ is independent of $G_k \sim N(0,1)$.

Turning now to **Step 1** in Section 4.4, we seek a fixed point $(\hat{r}^*, \hat{\beta}^*, \bar{b}_* > 0, \tilde{\mu}_*, \tilde{\sigma}_*, \tau_*)$ of (4.53)–(4.54) satisfying

$$0 = \delta X^\top \mathrm{S}_{\bar{b}_*}(\hat{r}^*), \qquad \hat{r}^* = y - X\hat{\beta}^* + \mathrm{S}_{\bar{b}_*}(\hat{r}^*), \tag{4.55}$$

$$\tilde{\mu}_* = \delta\,\mathbb{E}\big(\mathrm{S}'_{\bar{b}_*}(\bar{\varepsilon} + \tau_* G_*)\big) = 1, \qquad \tau_*^2 = \delta\,\mathbb{E}\big(\mathrm{S}_{\bar{b}_*}(\bar{\varepsilon} + \tau_* G_*)^2\big), \qquad \tilde{\sigma}_* = \sqrt{\delta}\tau_*, \tag{4.56}$$

where $\bar{\varepsilon} \sim P_{\bar{\varepsilon}}$ is independent of $G_* \sim N(0,1)$, and $\mu_* = \tilde{\mu}_*/(\delta \bar{b}_*)$ and $\sigma_* = \tilde{\sigma}_*/(\delta \bar{b}_*)$ are fixed points of the original state evolution equation (4.6). By Proposition 4.4, $\hat{\beta}^*$ solves the M-estimation problem in (4.50). Assuming that

$$\mathrm{M} \text{ is continuously differentiable and } \mathrm{S} = \mathrm{M}' \text{ is absolutely continuous with } \sup_{w \in \mathbb{R}} \mathrm{S}'(w) < \infty, \tag{4.57}$$

Donoho and Montanari (2016, Lemma 6.5) showed that for any $\tau > 0$, the map $b \mapsto \mathbb{E}\big(\mathrm{S}'_b(\bar{\varepsilon} + \tau G_*)\big) =: F_\tau(b)$ is continuous on $(0, \infty)$ with $\lim_{b \to 0} F_\tau(b) = 0$ and $\lim_{b \to \infty} F_\tau(b) = 1$, and hence that there exists $b \equiv b_\tau > 0$ satisfying $\mathbb{E}\big(\mathrm{S}'_b(\bar{\varepsilon} + \tau G_*)\big) = \delta^{-1}$ for $\delta \in (1, \infty)$. Using this, they deduced that under (4.57), there exists a unique solution $(\tau_*, \bar{b}_*)$ to (4.56) for any such δ (Donoho and Montanari, 2016, Corollary 4.4).

The functions f_*, g_* in **Step 2** in Section 4.4 are given by $f_* \colon w \mapsto \delta \bar{b}_* w$ and $g_* \colon (u,v) \mapsto \mathrm{S}_{\bar{b}_*}(v-u)/\bar{b}_*$, so for $n \in \mathbb{N}$ and $p \equiv p_n$, the 'stationary' AMP iteration (4.35) can be written as

$$\hat{\beta}^{k+1} = X^\top \mathrm{S}_{\bar{b}_*}(\hat{r}^k) + \hat{\beta}^k, \qquad \hat{r}^{k+1} = y - X\hat{\beta}^{k+1} + \mathrm{S}_{\bar{b}_*}(\hat{r}^k) \qquad \text{for } k \in \mathbb{N}_0. \tag{4.58}$$

Here, $\hat{r}^0 = y - X\hat{\beta}^0$ and $\hat{\beta}^0 = \beta + \tilde{\sigma}_* \xi = f_*(\mu_* \beta + \sigma_* \xi) \in \mathbb{R}^p$, where $\xi \sim N_p(0, I_p)$ is independent of the signal $\beta \in \mathbb{R}^p$. This choice of oracle initialiser ensures that the corresponding state evolution sequence is stationary with $\tau_k = \tau_*$ for all $k \in \mathbb{N}_0$. Then under the conditions (G0)–(G5) of Theorem 4.2 with $r = 2$, it follows from Remark 4.3 and (4.10) that for each fixed $k \in \mathbb{N}$, the empirical distributions of the components of $\hat{r}^k - \varepsilon \in \mathbb{R}^n$ and $\hat{\beta}^k - \beta \in \mathbb{R}^p$ converge completely in d_2 to $N(0, \tau_*^2)$ and $N(0, \tilde{\sigma}_*^2) = N(0, \delta \tau_*^2)$ respectively as $n, p \to \infty$ with $n/p \to \delta \in (1, \infty)$.

We remark that this result can in fact be derived by directly transforming (4.58) into an abstract asymmetric AMP iteration of the form (2.10). Note in particular that since $h(z,v) = z + v$ in (4.1) for the linear model and S_η is 1-Lipschitz for all $\eta > 0$, the function $\tilde{g}_k = \tilde{g}_* \colon (z,u,v) \mapsto \mathrm{S}_{\bar{b}_*}(z + v - u)/\bar{b}_*$ in (G4) is indeed Lipschitz.

As in Section 4.5, the remaining ingredient (**Step 3** in Section 4.4) is to show that the iterates $\hat{\beta}^k$ in (4.58) converge in the sense of (4.36) to some $\hat{\beta}^*$ satisfying (4.55), which is an M-estimator by Proposition 4.4. Under (4.57) and the additional assumption that M is strongly convex, i.e. $\inf_{w\in\mathbb{R}}\mathrm{S}'(w)>0$, the conclusion of Donoho and Montanari (2016, Theorem 4.1) is indeed that

$$\lim_{k\to\infty}\operatorname*{c-lim}_{p\to\infty}\frac{\|\hat{\beta}^k-\hat{\beta}^*\|^2}{p}=0. \tag{4.59}$$

Together with the state evolution characterisation of the iterates in (4.58), this leads to the following characterisation of the asymptotic performance of the M-estimator.

Theorem 4.7 (Donoho and Montanari, 2016, Theorem 4.2). *Consider a sequence of linear models $y=X\beta+\varepsilon$ satisfying (G0) and (G1), with $n/p\to\delta\in(1,\infty)$ as $n,p\to\infty$. Assume that the loss function M is continuously differentiable, and that the score function $\mathrm{S}=\mathrm{M}'$ is absolutely continuous with $0<\inf_{w\in\mathbb{R}}\mathrm{S}'(w)\leq\sup_{w\in\mathbb{R}}\mathrm{S}'(w)<\infty$. Let $(\tau_*,\bar{b}_*)$ be the unique fixed point of (4.56). Then*

$$\sup_{\psi\in\mathrm{PL}_2(2,1)}\left|\frac{1}{p}\sum_{j=1}^p\psi(\hat{\beta}_j^{\mathrm{M}}-\beta_j,\beta_j)-\mathbb{E}\big(\psi(\sqrt{\delta}\tau_*G,\bar{\beta})\big)\right|\xrightarrow{c}0 \tag{4.60}$$

as $n,p\to\infty$ with $n/p\to\delta$, where $G\sim N(0,1)$. In particular, the asymptotic mean squared error of $\hat{\beta}^{\mathrm{M}}$ is given by

$$\operatorname*{c-lim}_{p\to\infty}\frac{\|\hat{\beta}^{\mathrm{M}}-\beta\|^2}{p}=V(\mathrm{S}_{\bar{b}_*};\bar{\varepsilon}+\tau_*G)=\frac{\mathbb{E}\big(\mathrm{S}_{\bar{b}_*}(\bar{\varepsilon}+\tau_*G)^2\big)}{\mathbb{E}\big(\mathrm{S}_{\bar{b}_*}'(\bar{\varepsilon}+\tau_*G)\big)^2}=\frac{\tau_*^2/\delta}{1/\delta^2}=\delta\tau_*^2. \tag{4.61}$$

Under condition (G1) on the signal vectors $\beta\in\mathbb{R}^p$, Theorem 4.7 provides the limiting joint empirical distribution of the entries of $\hat{\beta}^{\mathrm{M}},\beta\in\mathbb{R}^p$. It turns out that even in the absence of (G1), we have

$$\sup_{\psi\in\mathrm{PL}_1(2,1)}\left|\frac{1}{p}\sum_{j=1}^p\psi(\hat{\beta}_j^{\mathrm{M}}-\beta_j)-\mathbb{E}\big(\psi(\sqrt{\delta}\tau_*G)\big)\right|\xrightarrow{c}0,$$

as evidenced by the fact that $\bar{\beta}$ does not appear in the state evolution recursion (4.56). Comparing the variance functional $V(\mathrm{S}_{\bar{b}_*};\bar{\varepsilon}+\tau_*G)$ in (4.61) with that in the classical setting, namely $V(\mathrm{S};\bar{\varepsilon})$ in (4.52), we emphasise the following points of difference. First, the asymptotic variance in the high-dimensional setting depends on $\mathrm{S}_{\bar{b}_*}=\mathrm{M}_{\bar{b}_*}'$, the score function of a regularised version of M (rather than M itself). In addition, the 'effective noise' in the high-dimensional regime is $\bar{\varepsilon}+\tau_*G$, rather than $\bar{\varepsilon}$. In fact, Donoho and Montanari (2016, Corollary 4.3) showed that

$$V(\mathrm{S}_{\bar{b}_*};\bar{\varepsilon}+\tau_*G)\geq\frac{1}{1-\delta^{-1}}\cdot\frac{1}{I(P_{\bar{\varepsilon}})}, \tag{4.62}$$

where $I(P_\varepsilon)^{-1}$ is the classical lower bound. This shows that the M-estimator is inefficient in high dimensions, particularly so when δ is close to 1.

We also mention that Donoho and Montanari (2015, Theorem 2.2) extended the conclusion (4.60) to M-estimators defined with respect to the Huber loss function, which is not strongly convex on $\mathbb{R}$ and hence is not covered by Theorem 4.7. Donoho and Montanari (2016, Section 6) noted an interesting connection between the Lasso and Huber M-estimators, as a special case ($J\colon w\mapsto\lambda|w|$) of a duality relationship between the following optimisation problems:

(i) The regularised least squares problem

$$\text{minimise}\quad\frac{1}{2}\|\breve{y}-\breve{X}\breve{\beta}\|^2+\sum_{j=1}^p J(\breve{\beta}_j)\quad\text{over }\breve{\beta}\in\mathbb{R}^n$$

based on $\breve{X}\in\mathbb{R}^{(n-p)\times n}$ and $\breve{y}\in\mathbb{R}^{n-p}$, with convex penalty $J\colon\mathbb{R}\to\mathbb{R}$;

(ii) The unpenalised M-estimation problem (4.50) based on $X\in\mathbb{R}^{n\times p}$, $y\in\mathbb{R}^n$ satisfying $\breve{X}X=0$ and $\breve{y}=\breve{X}y$, with convex loss function $\mathrm{M}\colon w\mapsto\min_{z\in\mathbb{R}}\{J(z)+(z-w)^2/2\}$.

4.7 GAMP for logistic regression

To further illustrate the generality and utility of the GAMP framework, we will now demonstrate how it can be applied to a popular non-linear GLM, namely the logistic regression model with canonical logit link. Suppose that we observe $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^p \times \{0,1\}$ with

$$\mathbb{P}(y_i = 1 \,|\, x_i^\top \beta) = \frac{e^{x_i^\top \beta}}{1 + e^{x_i^\top \beta}} = \zeta'(x_i^\top \beta), \quad \text{where } \ \zeta(z) := \log(1 + e^z) \tag{4.63}$$

for $1 \leq i \leq n$. Equivalently, we may view this as an instance of the model (4.1) with $\varepsilon_1, \dots, \varepsilon_n \overset{\mathrm{iid}}{\sim} U[0,1]$ and $h(z,v) = \mathbb{1}_{\{v \leq \zeta'(z)\}}$, so that $y_i = h(x_i^\top \beta, \varepsilon_i) = \mathbb{1}_{\{\varepsilon_i \leq \zeta'(x_i^\top \beta)\}}$ for each i, and seek to estimate $\beta \in \mathbb{R}^p$ by maximum likelihood via

$$\hat{\beta}^{\mathrm{MLE}} \in \operatorname*{argmin}_{\tilde{\beta} \in \mathbb{R}^p} \sum_{i=1}^n \bigl\{\zeta(x_i^\top \tilde{\beta}) - y_i x_i^\top \tilde{\beta}\bigr\}, \tag{4.64}$$

where the objective function in (4.64) is the negative log-likelihood. Albert and Anderson (1984) showed that this MLE exists if and only if $\mathcal{X}_0 := \{x_i : 1 \leq i \leq n,\, y_i = 0\}$ and $\mathcal{X}_1 := \{x_i : 1 \leq i \leq n,\, y_i = 1\}$ are not (strongly) linearly separable, i.e. for any $\tilde{\beta} \neq 0$, there either exists $x_{i_0} \in \mathcal{X}_0$ with $x_{i_0}^\top \tilde{\beta} > 0$ or $x_{i_1} \in \mathcal{X}_1$ with $x_{i_1}^\top \tilde{\beta} < 0$. In the random design setting of (G0) where $x_1, \dots, x_n \overset{\mathrm{iid}}{\sim} N_p(0, I_p/n)$ for each n and $p \equiv p_n$, Candès and Sur (2020) established a sharp phase transition for the existence of $\hat{\beta}^{\mathrm{MLE}}$. Specifically, they proved that there exists a decreasing function $s_{\mathrm{MLE}} \colon (0,\infty) \to [0,\infty)$ with the following property: if the signals $\beta \in \mathbb{R}^p$ are such that $n^{-1/2}\|\beta\| \xrightarrow{c} \kappa \in (0,\infty)$ as $n, p \to \infty$ with $n/p \to \delta \in (1,\infty)$, then $\hat{\beta}^{\mathrm{MLE}}$ exists with probability tending to 0 if $\kappa > s_{\mathrm{MLE}}(1/\delta)$, and exists with probability tending to 1 if $\kappa < s_{\mathrm{MLE}}(1/\delta)$.

Henceforth, we will restrict attention to the latter regime, and use the GAMP formalism in Sections 4.1 and 4.4 to explain how to derive a result of Sur and Candès (2019a,b) on the high-dimensional asymptotics of $\hat{\beta}^{\mathrm{MLE}}$, which is formally stated as Theorem 4.8 below. Recall from (4.5) that for a sequence of logistic regression models (4.63) satisfying (G1), the asymptotic signal strength $\kappa^2 = \text{c-lim}_{n\to\infty} \|\beta\|^2/n$ is equal to $\mathbb{E}(\bar{\beta}^2)/\delta$. Noting that $\hat{\beta}^{\mathrm{MLE}}$ in (4.64) solves a convex optimisation problem of the form (4.22) with $J \equiv 0$ and $\ell(u,v) = \zeta(u) - vu$, we see that the functions $\bar{g}_k, g_k, f_{k+1}$ in (4.23)–(4.24) are given by

$$\begin{aligned} \bar{g}_k(u,v) &= \operatorname{prox}_{\bar{b}_k \zeta}(u + \bar{b}_k v) = u + \bar{b}_k v - \bar{b}_k \zeta'\bigl(\operatorname{prox}_{\bar{b}_k \zeta}(u + \bar{b}_k v)\bigr), \\ g_k(u,v) &= v - \zeta'\bigl(\operatorname{prox}_{\bar{b}_k \zeta}(u + \bar{b}_k v)\bigr), \qquad f_{k+1}(w) = -\frac{w}{\bar{c}_k} \end{aligned} \tag{4.65}$$

for $k \in \mathbb{N}_0$, since $b\zeta'(\operatorname{prox}_{b\zeta}(u)) + \operatorname{prox}_{b\zeta}(u) - u = 0$ by the definition of $\operatorname{prox}_{b\zeta}$ in (4.25) for $b > 0$.

Given $\bar{b}_0 > 0$, $\hat{\beta}^0 \in \mathbb{R}^p$ and $\theta^0 := X\hat{\beta}^0$, the GAMP recursion (4.29) therefore takes the form

$$\begin{aligned} \hat{\beta}^{k+1} &= \delta \bar{b}_{k+1} X^\top \bigl\{y - \zeta'\bigl(\operatorname{prox}_{\bar{b}_k \zeta}(\theta^k + \bar{b}_k y)\bigr)\bigr\} + \frac{\bar{b}_{k+1}}{\bar{b}_k} \hat{\beta}^k, \\ \theta^{k+1} &= X\hat{\beta}^{k+1} - \bar{b}_{k+1}\bigl\{y - \zeta'\bigl(\operatorname{prox}_{\bar{b}_k \zeta}(\theta^k + \bar{b}_k y)\bigr)\bigr\} \end{aligned} \tag{4.66}$$

for $k \in \mathbb{N}_0$, where $\bar{b}_{k+1} = -1/(\delta \bar{c}_k)$. Using (4.9) from Lemma 4.1, as well as (4.27), we now write the corresponding state evolution recursion (4.6)–(4.7) for (4.66) in terms of $\tilde{\mu}_k := \delta \bar{b}_k \mu_k$ and $\tilde{\sigma}_k := \delta \bar{b}_k \sigma_k$. This yields

$$\begin{aligned} \bar{b}_{k+1} &= \frac{\bar{b}_k}{\delta}\left(1 - \mathbb{E}\left\{\frac{1}{1 + \bar{b}_k \zeta''\bigl(\operatorname{prox}_{\bar{b}_k \zeta}(Z_k + \bar{b}_k Y)\bigr)}\right\}\right)^{-1}, \\ \tilde{\mu}_{k+1} &= \frac{\delta^2 \bar{b}_{k+1}}{\mathbb{E}(\bar{\beta}^2)} \mathbb{E}\Bigl(Z\Bigl\{Y - \zeta'\bigl(\operatorname{prox}_{\bar{b}_k \zeta}(Z_k + \bar{b}_k Y)\bigr)\Bigr\}\Bigr) + \tilde{\mu}_k, \\ \tilde{\sigma}_{k+1}^2 &= \delta^2 \bar{b}_{k+1}^2 \, \mathbb{E}\Bigl(\Bigl\{Y - \zeta'\bigl(\operatorname{prox}_{\bar{b}_k \zeta}(Z_k + \bar{b}_k Y)\bigr)\Bigr\}^2\Bigr) \end{aligned} \tag{4.67}$$

for $k \in \mathbb{N}_0$, where given independent $Z \sim N(0, \mathbb{E}(\bar{\beta}^2)/\delta)$, $\tilde{G}_k \sim N(0,1)$ and $\bar{\varepsilon} \sim P_{\bar{\varepsilon}}$, we set

$$Y = h(Z, \bar{\varepsilon}) = \mathbb{1}_{\{\bar{\varepsilon} \leq \zeta'(Z)\}}, \qquad Z_k = \mu_{Z,k} Z + \sigma_{Z,k} \tilde{G}_k = \tilde{\mu}_k Z + \delta^{-1/2}\, \tilde{\sigma}_k \tilde{G}_k$$

in view of (4.63), (4.8) and the definition of f_{k+1} in (4.65). Sur and Candès (2019b, Section 3.1) showed that (4.67) is equivalent to the original state evolution recursion they defined in Sur and Candès (2019a, Section 4.1).

In accordance with **Step 1** in Section 4.4, we seek a fixed point $(\hat{\beta}^*, \theta^*, \tilde{\mu}_*, \tilde{\sigma}_*, \bar{b}_* > 0)$ of (4.66)–(4.67) satisfying

$$\theta^* = X\hat{\beta}^* - \bar{b}_*\bigl\{y - \zeta'\bigl(\mathrm{prox}_{\bar{b}_*\zeta}(\theta^* + \bar{b}_* y)\bigr)\bigr\}, \qquad 0 = X^\top\bigl\{y - \zeta'\bigl(\mathrm{prox}_{\bar{b}_*\zeta}(\theta^* + \bar{b}_* y)\bigr)\bigr\}, \tag{4.68}$$

$$\tilde{\sigma}_*^2 = \delta^2 \bar{b}_*^2\, \mathbb{E}\Bigl(\Bigl\{Y - \zeta'\bigl(\mathrm{prox}_{\bar{b}_*\zeta}(Z_* + \bar{b}_* Y)\bigr)\Bigr\}^2\Bigr), \qquad 0 = \mathbb{E}\Bigl(Z\Bigl\{Y - \zeta'\bigl(\mathrm{prox}_{\bar{b}_*\zeta}(Z_* + \bar{b}_* Y)\bigr)\Bigr\}\Bigr), \tag{4.69}$$

$$1 - \frac{1}{\delta} = \mathbb{E}\left\{\frac{1}{1 + \bar{b}_*\zeta''\bigl(\mathrm{prox}_{\bar{b}_*\zeta}(Z_* + \bar{b}_* Y)\bigr)}\right\}, \tag{4.70}$$

where $Z_* := \tilde{\mu}_* Z + \delta^{-1/2}\, \tilde{\sigma}_* \tilde{G}_*$ with $Z \sim N(0, \mathbb{E}(\bar{\beta}^2)/\delta)$ independent of $\tilde{G}_* \sim N(0,1)$. It turns out that there exists a unique solution $(\tilde{\mu}_*, \tilde{\sigma}_*, \bar{b}_* > 0)$ to (4.69)–(4.70) precisely when $\mathbb{E}(\bar{\beta}^2)/\delta \equiv \kappa^2 < s_{\mathrm{MLE}}(1/\delta)^2$ (Sur and Candès, 2019b, Lemma 7 and Remark 1), in which case $\hat{\beta}^{\mathrm{MLE}}$ exists with probability tending to 1. By Proposition 4.4, $\hat{\beta}^*$ in (4.68) is an MLE for β in the logistic regression model.

Proceeding as in **Step 2** in Section 4.4, we can use the fixed points in (4.68)–(4.70) to construct a stationary version of (4.66) based on $f_*\colon w \mapsto \delta\bar{b}_* w$ and $g_*\colon (u,v) \mapsto v - \zeta'\bigl(\mathrm{prox}_{\bar{b}_*\zeta}(u + \bar{b}_* v)\bigr)$. For each $n \in \mathbb{N}$ and $p \equiv p_n$, let $\hat{\beta}^0 := \tilde{\mu}_*\beta + \tilde{\sigma}_*\xi = f_*(\mu_*\beta + \sigma_*\xi) \in \mathbb{R}^p$ be an oracle initialiser with $\xi \sim N_p(0, I_p)$ taken to be independent of the signal $\beta \in \mathbb{R}^p$. Then setting $\theta^0 = X\hat{\beta}^0$, we inductively define

$$\hat{\beta}^{k+1} = \delta\bar{b}_* X^\top\bigl\{y - \zeta'\bigl(\mathrm{prox}_{\bar{b}_*\zeta}(\theta^k + \bar{b}_* y)\bigr)\bigr\} + \hat{\beta}^k, \qquad \theta^{k+1} = X\hat{\beta}^k - \bar{b}_*\bigl\{y - \zeta'\bigl(\mathrm{prox}_{\bar{b}_*\zeta}(\theta^k + \bar{b}_* y)\bigr)\bigr\} \tag{4.71}$$

for $k \in \mathbb{N}_0$. By the choice of $\hat{\beta}^0$ above, the associated state evolution recursion (4.67) is stationary, i.e. $\tilde{\mu}_k = \tilde{\mu}_*$ and $\tilde{\sigma}_k = \tilde{\sigma}_*$ for all $k \in \mathbb{N}_0$. Consequently, under the hypotheses of Theorem 4.2 with $r = 2$, it follows from Remark 4.3 that for each fixed $k \in \mathbb{N}$, the joint empirical distribution of the entries of $\hat{\beta}^k, \beta \in \mathbb{R}^p$ converges completely in d_2 to the distribution of $(\tilde{\mu}_*\bar{\beta} + \tilde{\sigma}_* G, \bar{\beta})$ as $n, p \to \infty$ with $n/p \to \delta$, where $\bar{\beta} \sim \pi_{\bar{\beta}}$ is independent of $G \sim N(0,1)$. On a technical note, we remark that the function

$$\tilde{g}_*\colon (z,u,v) \mapsto g_*(u, h(z,v)) = \mathbb{1}_{\{v \leq \zeta'(z)\}} - \zeta'\Bigl(\mathrm{prox}_{\bar{b}_*\zeta}\bigl(u + \bar{b}_* \mathbb{1}_{\{v \leq \zeta'(z)\}}\bigr)\Bigr)$$

in (G4) is not Lipschitz since $h\colon (z,v) \mapsto \mathbb{1}_{\{v \leq \zeta'(z)\}}$ is not continuous, so an additional approximation argument is needed to formally justify the application of Theorem 4.2.

Finally, we discuss **Step 3** in Section 4.4, whose aim is to show that the iterates in (4.71) converge in the sense of (4.36) to a fixed point $\hat{\beta}^* \equiv \hat{\beta}^{\mathrm{MLE}}$ satisfying (4.68). This is the content of Sur and Candès (2019b, Theorem 7), and follows from similar arguments to those used by Donoho and Montanari (2016) to prove (4.59) for the M-estimators in Section 4.6. An additional technical obstacle in this setting is that $\zeta\colon z \mapsto \log(1 + e^z)$ and hence the negative log-likelihood function in (4.64) are strongly convex on compact sets but not on the entirety of their domains. One way to address this issue is to show that $\hat{\beta}^k, \hat{\beta}^{\mathrm{MLE}}$ are contained in some sufficiently large Euclidean ball with overwhelming probability. Indeed, it follows from the state evolution characterisation of (4.71) that $\|\hat{\beta}^k\|^2/p = O_c(1)$ for each fixed k; in addition, Sur and Candès (2019b, Theorem 4) established the boundedness property $\|\hat{\beta}^{\mathrm{MLE}}\|^2/p = O_c(1)$ in the regime $\kappa < s_{\mathrm{MLE}}(1/\delta)$ where $\hat{\beta}^{\mathrm{MLE}}$ exists with probability tending to 1.

Theorem 4.8 (Sur and Candès, 2019a, Theorem 2)**.** *Consider a sequence of logistic regression models* (4.63) *satisfying* (G0) *and* (G1) *for $r = 2$ as $n, p \to \infty$ with $n/p \to \delta \in (1, \infty)$. Assume that*

$\mathbb{E}(\bar{\beta}^2)/\delta \equiv \kappa^2 < s_{\mathrm{MLE}}(1/\delta)^2$, so that (4.64) defines a maximum likelihood estimator $\hat{\beta}^{\mathrm{MLE}}$ with probability tending to 1, and there exist $\tilde{\mu}_*, \tilde{\sigma}_*, \bar{b}_*$ satisfying (4.69)–(4.70). Then

$$\sup_{\psi \in \mathrm{PL}_2(2,1)} \left| \frac{1}{p} \sum_{j=1}^{p} \psi\big(\hat{\beta}_j^{\mathrm{MLE}} - \tilde{\mu}_* \beta_j, \beta_j\big) - \mathbb{E}\big(\psi(\tilde{\sigma}_* G, \bar{\beta})\big) \right| \xrightarrow{c} 0$$

as $n, p \to \infty$ with $n/p \to \delta$, where $G \sim N(0,1)$ is independent of $\bar{\beta} \sim \pi_{\bar{\beta}}$. In particular,

$$\frac{1}{p} \sum_{j=1}^{p} \big(\hat{\beta}_j^{\mathrm{MLE}} - \tilde{\mu}_* \beta_j\big) \xrightarrow{c} 0, \quad \frac{1}{p} \sum_{j=1}^{p} \big(\hat{\beta}_j^{\mathrm{MLE}} - \tilde{\mu}_* \beta_j\big)^2 \xrightarrow{c} \tilde{\sigma}_*^2, \quad \frac{\|\hat{\beta}^{\mathrm{MLE}} - \beta\|^2}{p} \xrightarrow{c} (\tilde{\mu}_* - 1)^2\, \mathbb{E}(\bar{\beta}^2) + \tilde{\sigma}_*^2.$$

Thus, for large p, the components of $\hat{\beta}^{\mathrm{MLE}} \in \mathbb{R}^p$ have approximately the same empirical distribution as those of $\tilde{\mu}_*\beta + \tilde{\sigma}_*\xi$ (the oracle initialiser $\hat{\beta}^0$ in (4.71) above), so we can interpret $\tilde{\mu}_*$ as an asymptotic bias factor and $\tilde{\sigma}_*^2$ as a limiting variance. Sur and Candès (2019a) observe empirically that when $n, p \to \infty$ with $\delta \in (1, \infty)$, both the limiting bias and variance are larger than they would be in classical settings where p is fixed or grows sufficiently slowly with n (in which case $\hat{\beta}^{\mathrm{MLE}}$ would be asymptotically unbiased (with $\tilde{\mu}_* = 1$) and asymptotically efficient as $n \to \infty$). Their Figure 7 illustrates that this high-dimensional phenomenon becomes increasingly pronounced when either δ is reduced or κ is enlarged; in fact, when κ approaches the critical value $s_{\mathrm{MLE}}(1/\delta)$ for the existence of $\hat{\beta}^{\mathrm{MLE}}$, the value of $\tilde{\mu}_*$ diverges to infinity, as does the ratio between $\tilde{\sigma}_*$ and the Cramér–Rao lower bound.

It is instructive to compare the high-dimensional asymptotic performance of $\hat{\beta}^{\mathrm{MLE}}$ in the logistic model with that of the M-estimator (4.50) in the linear model. Note that while both estimators exhibit variance inflation (as quantified by Theorems 4.7 and 4.8), only the former suffers from bias inflation. Indeed, in the linear model, the AMP state evolution recursion (4.56) yields $\mu_k = 1$ for all k, and hence $\mu_* = 1$ (implicitly) in Theorem 4.7 for the M-estimator; see also (4.20) in Section 4.3.

5 Conclusions and extensions

With the abstract AMP recursions in Section 2 as our starting point, we have shown how to design and analyse AMP algorithms for estimating structured signals, both in low-rank spiked models with Gaussian noise matrices and in GLMs with Gaussian design matrices. In high-dimensional asymptotic regimes where the matrix dimensions scale proportionally to each other, we have illustrated how to apply the abstract master theorems to derive precise state evolution characterisations of AMP estimation performance, which we have stated as complete convergence guarantees.

In Section 4, we have presented a general recipe that uses AMP systematically to obtain exact expressions for the asymptotic error of penalised and unpenalised M-estimators in GLMs with Gaussian design matrices. An alternative approach to deriving such guarantees is via Gaussian comparison inequalities and the convex Gaussian min-max theorem (CGMT) (Thrampoulidis et al., 2015). These techniques have recently been used to analyse the performance of regularised M-estimators (Thrampoulidis et al., 2018), the Lasso (Miolane and Montanari, 2021), boosting (Liang and Sur, 2022) and convex-constrained least squares estimators (Han, 2022), as well as to elucidate the so-called 'double descent' phenomenon in over-parametrised binary classification models (Deng et al., 2019; Kini and Thrampoulidis, 2020).

Remaining within the realm of Gaussian matrices, we mention that the results in this monograph can be extended to AMP recursions with (i) non-separable denoising functions that do not act componentwise on their vector arguments, and can therefore take advantage of correlation between entries of the signal (Ma et al., 2019; Berthier et al., 2020); (ii) matrices with independent entries and a blockwise variance structure (Javanmard and Montanari, 2013). With a carefully chosen variance structure ('spatial coupling'), AMP has been shown to achieve the information-theoretic limit for compressed sensing (Donoho et al., 2013).

In the setting of AMP for asymmetric matrices in Section 2.2, the results of Theorem 2.5 can be generalised to matrices with i.i.d. sub-Gaussian entries with mild additional assumptions (Bayati et al., 2015; Chen and Lam, 2021). It is likely that the proof strategies in these papers can be developed further to extend other theoretical results (such as Theorem 4.2 for GAMP) to these more general random matrix ensembles.

When the data matrix does not have i.i.d. Gaussian entries, AMP is not guaranteed to converge, and in fact can even diverge in sometimes pathological ways; see Rangan et al. (2019a) for a discussion of this issue. For this reason, a number of other AMP-based algorithms have been introduced that allow for this assumption to be weakened in various ways, such as Vector AMP (VAMP) (Rangan et al., 2019b), orthogonal AMP (OAMP) (Ma and Ping, 2017; Takeuchi, 2020) and other generalisations of AMP for rotationally invariant matrices (Opper et al., 2016; Fan, 2022).

Vector AMP (VAMP) is an iterative algorithm (based on Expectation Propagation) for estimation in rotationally invariant linear models (Rangan et al., 2019b; Takeuchi, 2020, 2021b) and generalised linear models (Schniter et al., 2016; Pandit et al., 2020). Rangan et al. (2019b) and Pandit et al. (2020) showed that the asymptotic estimation error of VAMP (with optimal denoising functions) coincides with the statistical physics-based prediction for the Bayes-optimal error whenever the state evolution recursion has a unique fixed point. Orthogonal AMP (Ma and Ping, 2017; Takeuchi, 2020) is an algorithm that is equivalent to VAMP for estimation in rotationally invariant linear models. Recently, Ma et al. (2021) studied the performance of Expectation Propagation (an algorithm closely related to VAMP) for rotationally invariant GLMs, and analysed the impact of the spectrum on the estimation performance. VAMP has also been used to obtain the asymptotic risk of convex-penalized estimators for rotationally invariant GLMs (Gerbelot et al., 2020a,b). A few lower complexity alternatives to VAMP have also been proposed, including convolutional AMP (Takeuchi, 2021a), Memory AMP for linear models (Liu et al., 2021), and Generalised Memory AMP for GLMs (Tian et al., 2021).

Fan (2022) and Zhong et al. (2021) provide a master theorem for an abstract AMP recursion defined via a rotationally invariant random matrix. Compared with the original Gaussian setting of Section 2, AMP recursions with general rotationally invariant random matrices have two key differences, as seen in (3.38): (i) the presence of multiple memory terms, accounting for all the preceding iterates, and (ii) thresholding functions that act on all the preceding iterates rather than just the current one. The abstract AMP recursion of Fan (2022) and Zhong et al. (2021) has been used to derive AMP algorithms with state evolution guarantees for low-rank matrix estimation with rotationally invariant noise (Opper et al., 2016; Çakmak and Opper, 2019; Fan, 2022; Zhong et al., 2021; Mondelli and Venkataramanan, 2021), as discussed in Section 3.5. See also Venkataramanan et al. (2021) for an application of this more general AMP framework to GLMs defined via rotationally invariant matrices.

Though the focus in this tutorial has been on low-rank matrix estimation and generalised linear models, both AMP and Vector AMP have been applied to a number of other statistical problems including tensor PCA (Montanari and Richard, 2014) and inference in multilayer neural networks (Manoel et al., 2017; Fletcher et al., 2018; Emami et al., 2020). AMP has also been used to obtain lower bounds on the limiting estimation error of a broad class of general first-order methods such as gradient descent and mirror descent (Celentano et al., 2020). An active area of current research is to determine whether AMP outperforms all other polynomial-time algorithms in low-rank matrix estimation and GLMs. In these settings, the statistical-computational gap has been precisely characterised in terms of the critical points of a 'potential function' (Lelarge and Miolane, 2019; Barbier et al., 2019). As mentioned in Section 3.3, the performance of both Bayes-AMP and the Bayes optimal estimator correspond to (possibly different) critical points of this function, and when the potential function has a single critical point, Bayes-AMP achieves Bayes optimal performance. This connection suggests that AMP will play an important role in understanding statistical-computational gaps in a wider context.

6 Appendix: proofs and technical remarks

In addition to the definitions in Section 1.1, we introduce the following notation. The Moore–Penrose pseudoinverse of a matrix $A \in \mathbb{R}^{k\times\ell}$ will be denoted by $A^+ \in \mathbb{R}^{\ell\times k}$. This satisfies $A^+ = (A^\top A)^+ A^\top$ (e.g. Barata and Hussein, 2012, Proposition 3.2), and if $k = \ell$ and A is invertible, then $A^+ = A^{-1}$. For non-negative, real-valued functions f, g, we write $f \lesssim g$ if there exists a universal constant $C > 0$ such that $f \leq Cg$; more generally, given parameters $\alpha_1, \dots, \alpha_N$, we write $f \lesssim_{\alpha_1,\dots,\alpha_N} g$ if there exists $C \equiv C_{\alpha_1,\dots,\alpha_N} > 0$, depending only on $\alpha_1, \dots, \alpha_N$, such that $f \leq Cg$.

6.1 Technical remarks on the master theorems in Section 2.1

In this subsection, we will make some general observations that unify Theorems 2.1 and 2.3 with other master theorems in the AMP literature (e.g. Bolthausen, 2014; Bayati and Montanari, 2011; Javanmard and Montanari, 2013). There are a number of respects in which our results are presented differently and/or in slightly greater generality, and we discuss each of these in turn.

Remark 6.1 (*Complete convergence*)**.** In Section 6.4, we will also establish the following variants of Theorem 2.1, neither of which implies the other (or the original theorem): for a sequence of symmetric AMP recursions (2.1) satisfying (A0), (A4) and (A5), and an associated sequence of state evolution parameters $(\tau_k^2 : k \in \mathbb{N})$ as in (2.2), the following hold for each fixed $k \in \mathbb{N}$ as $n \to \infty$:

(a) Suppose that (A1)–(A3) hold with $\overset{p}{\to}$ and $O_p(1)$ in place of $\overset{c}{\to}$ and $O_c(1)$ respectively. Then $d_r\bigl(\nu_n(h^k, \gamma), N(0, \tau_k^2) \otimes \pi\bigr) \overset{p}{\to} 0$, or equivalently $\widetilde{d}_r\bigl(\nu_n(h^k, \gamma), N(0, \tau_k^2) \otimes \pi\bigr) \overset{p}{\to} 0$.

(b) Suppose instead that (A1)–(A3) hold with $\overset{a.s.}{\to}$ and $O_{a.s.}(1)$ in place of $\overset{c}{\to}$ and $O_c(1)$ respectively, and moreover that $\bigl(W(n) : n \in \mathbb{N}\bigr)$ is independent of $\bigl(m^0(n), \gamma(n) : n \in \mathbb{N}\bigr)$. Then $d_r\bigl(\nu_n(h^k, \gamma), N(0, \tau_k^2) \otimes \pi\bigr) \overset{a.s.}{\to} 0$, or equivalently $\widetilde{d}_r\bigl(\nu_n(h^k, \gamma), N(0, \tau_k^2) \otimes \pi\bigr) \overset{a.s.}{\to} 0$.

Stronger versions of these statements can be formulated as analogues of Theorem 2.3. We now explain why we have stated our AMP master theorems (and all subsequent asymptotic results in the monograph) in terms of complete convergence.

- Complete convergence is stronger than almost sure convergence and convergence in probability, so the conclusions of Theorems 2.1 and 2.3 provide stronger convergence guarantees than (a) and (b).

- In view of Remark 7.1, neither the conditions (A0)–(A3) nor their analogues in (a) impose any restrictions on the dependence structure across $n \in \mathbb{N}$ of the random triples $\bigl(m^0(n), \gamma(n), W(n)\bigr)$ that generate the AMP iterates. By contrast, the additional assumption in (b) is somewhat unnatural from a statistical point of view, except perhaps when $\bigl(m^0(n), \gamma(n) : n \in \mathbb{N}\bigr)$ is taken to be deterministic sequence that satisfies the other conditions in (b). Note however that this special case is covered by Theorems 2.1 and 2.3, which yield stronger conclusions than (b), as mentioned above.

- The method of proof of Theorems 2.1 and 2.3 (via Proposition 6.16) is well-suited to complete convergence and convergence in probability, but appears not to be able to handle almost sure convergence directly; it is not clear whether (b) holds in general if we only assume (A0) rather than the stronger independence condition above. The reason for this is that in many of the key technical arguments, the convergence of some random sequence (X_n) of interest is established by first identifying a more tractable sequence (Y_n) such that $Y_n \overset{d}{=} X_n$ for all n. To show that $X_n \overset{c}{\to} x$ for some deterministic x, or that $X_n = O_c(1)$, it suffices to prove that $Y_n \overset{c}{\to} x$ or $Y_n = O_c(1)$ respectively in view of Definition 1.1 of complete convergence. Similarly, $Y_n \overset{p}{\to} x$ implies that $X_n \overset{p}{\to} x$, and $Y_n = O_p(1)$ implies that $X_n = O_p(1)$. However, if $Y_n \overset{a.s.}{\to} x$, then it does not necessarily follow that $X_n \overset{a.s.}{\to} x$, and if $Y_n = O_{a.s.}(1)$, then it need not be the case that $X_n = O_{a.s.}(1)$.

Remark 6.2 (*Uniformity over* $\mathrm{PL}_D(r,1)$ *and the link between pseudo-Lipschitz functions and Wasserstein convergence*)**.** Many asymptotic convergence results for AMP iterations are stated in the form

$$\frac{1}{n}\sum_{i=1}^{n}\psi(X_{ni}^k) \rightsquigarrow \mathbb{E}\bigl(\psi(\bar{X}^k)\bigr) \in \mathbb{R} \;\text{ as } n\to\infty, \text{ for every } \psi\in\mathrm{PL}_D(r), \tag{6.1}$$

where $r\in[2,\infty)$, $\rightsquigarrow$ denotes one of the three modes of stochastic convergence discussed in Remark 6.1, the random vectors $\bar{X}^k, X_{ni}^k$ take values in $\mathbb{R}^D$ for some fixed $D\in\mathbb{N}$, and $k\in\mathbb{N}$ is a fixed iteration number; usually, each X_{ni}^k depends on the i^{th} coordinates of vector quantities in the first k iterations of an AMP recursion indexed by n. Recalling the definition (1.4) of $\widetilde{d}_r$, we deduce from Corollary 7.21 that any conclusion of the form (6.1) can be automatically upgraded to a uniform statement

$$\widetilde{d}_r(\mu_n^k,\bar{\mu}^k) = \sup_{\psi\in\mathrm{PL}_D(r,1)}\left|\frac{1}{n}\sum_{i=1}^{n}\psi(X_{ni}^k) - \mathbb{E}\bigl(\psi(\bar{X}^k)\bigr)\right| \rightsquigarrow 0 \;\text{ as } n\to\infty \tag{6.2}$$

featuring the same mode of convergence $\rightsquigarrow$ as in (6.1), where we write μ_n^k for the empirical distribution of $X_{n1}^k,\dots,X_{nn}^k$ on $\mathbb{R}^D$, and $\bar{\mu}^k$ for the distribution of the limiting random vector $\bar{X}^k$. Furthermore, by Corollary 7.21, both (6.1) and (6.2) are equivalent to the assertion that $d_r(\mu_n^k,\bar{\mu}^k)\rightsquigarrow 0$. In essence, this is because $d_r,\widetilde{d}_r$ are equivalent metrics, in the sense that they generate the same topology on the space $\mathcal{P}_D(r)$ of probability distributions on $\mathbb{R}^D$ with a finite r^{th} moment; see Theorem 7.17 and Remark 7.18.

On a technical note, the measurability of the random quantities $\widetilde{d}_r(\mu_n^k,\bar{\mu}^k)$ and $d_r(\mu_n^k,\bar{\mu}^k)$ is guaranteed by analytic considerations; it is shown in Proposition 7.16 that the supremum in (6.2) can instead be taken over a deterministic countable subset $T'\subseteq\mathrm{PL}_D(r)$ of bounded Lipschitz functions.

Remark 6.3 (*Finite-sample analysis*)**.** To complement and refine some of the asymptotic conclusions of the type (6.1) for general AMP procedures, the relevant proof techniques have been adapted to establish concentration inequalities for quantities of the form $n^{-1}\sum_{i=1}^n\psi(X_{ni}^k)-\mathbb{E}\bigl(\psi(\bar{X}^k)\bigr)$ for $k,n\in\mathbb{N}$ and fixed arbitrary $\psi\in\mathrm{PL}_D(r,1)$, under suitable assumptions. For $r=2$, such finite-sample guarantees were obtained for asymmetric recursions by Rush and Venkataramanan (2018) and for symmetric recursions by Barbier et al. (2020). Their conclusions can be generalised to $r>2$ with the aid of Lemma 7.12, a general concentration result for sums of pseudo-Lipschitz functions of independent Gaussian random variables. It would be interesting to see whether the above results can be extended to derive a stronger finite-sample analogue of Theorem 2.1 in the form of a concentration inequality for $\widetilde{d}_r\bigl(\nu_n(h^k,\gamma),N(0,\tau_k^2)\otimes\pi\bigr) = \sup_{\psi\in\mathrm{PL}_2(r,1)}\bigl|n^{-1}\sum_{i=1}^n\psi(h_i^k,\gamma_i)-\mathbb{E}\bigl(\psi(G_k,\bar{\gamma})\bigr)\bigr|$ or $d_r\bigl(\nu_n(h^k,\gamma),N(0,\tau_k^2)\otimes\pi\bigr)$ for $k,n\in\mathbb{N}$.

Remark 6.4 (*Conditions* (A2) *and* (A3))**.** For $r\geq 2$, conclusions of the form (6.1) have previously been derived for general AMP iterations under a boundedness assumption on the $(2r-2)^{th}$ moments of the empirical distributions $\nu_n(m^0)$ for $n\in\mathbb{N}$. In (A2), we relax this to a boundedness condition $\|m^0\|_{n,r}=O_c(1)$ on the empirical r^{th} moments, which is more natural and in line with what one would expect for a d_r convergence result. To accommodate this weaker assumption, we apply Hölder's inequality rather than the Cauchy–Schwarz inequality in Lemma 7.24, which is used in a key estimate in the proof of Proposition 6.16(c) below; see (6.30) and (6.47). By making similar alterations to the statements and proofs of other AMP results, it ought to be possible to avoid any mention of $(2r-2)^{th}$ empirical moments.

The primary purpose of (A3) is to ensure that the asymptotic dependence between different iterates h^j,h^ℓ (as measured by the inner product $\langle h^j,h^\ell\rangle_n$ between them) has a deterministic limiting expression, namely $\bar{\mathrm{T}}_{j,\ell}$ as defined in (2.6); see also Proposition 6.16(d, e, f). The existence of the limiting covariance structure captured by (2.6) is crucial to the success of the proof strategy for Theorems 2.1 and 2.3; in fact, its existence is a *necessary* condition for the more general conclusion in Theorem 2.3, as can be seen by taking $\psi(x_1,\dots,x_k):=x_jx_\ell$ therein for $1\leq j,\ell\leq k$.

Remark 6.5. Since $\pi \in \mathcal{P}_1(r)$ by (A1), recall from Section 1.1 that if $\bar{\gamma} \sim \pi$, then $\mathbb{E}\big(\psi(\bar{\gamma})\big) = \int_{\mathbb{R}} \psi\, d\pi < \infty$ for all $\psi \in \mathrm{PL}_1(r)$, the set of all pseudo-Lipschitz functions on $\mathbb{R}$ of order r. Thus, in (A3), given Lipschitz functions F_0, ϕ on $\mathbb{R}$, Lemma 7.22 ensures that $x \mapsto F_0(x)\phi(x)$ lies in $\mathrm{PL}_1(2) \subseteq \mathrm{PL}_1(r)$ since $r \geq 2$, so $\mathbb{E}\big(F_0(\bar{\gamma})\phi(\bar{\gamma})\big)$ is finite.

It can be shown by fairly routine arguments that the following condition implies the first condition in (A2) as well as (A3); see Section 6.6 for a full justification.

(A1$^+$) There exists a Lipschitz function $\tilde{f}_0 \colon \mathbb{R}^2 \to \mathbb{R}$ and a probability distribution $\tilde{\nu}^0 \in \mathcal{P}_1(2)$ such that writing μ^0 for the distribution of $\big(\tilde{f}_0(\bar{\eta}, \bar{\gamma}), \bar{\gamma}\big)$ when $\bar{\eta} \sim \tilde{\nu}^0$ and $\bar{\gamma} \sim \pi$ are independent, we have $d_2\big(\nu_n(m^0, \gamma), \mu^0\big) \xrightarrow{c} 0$.

In applications, (A1$^+$) can be more convenient to verify than (A3). Note that if $d_2\big(\nu_n(h^0, \gamma), \tilde{\nu}^0 \otimes \pi\big) \xrightarrow{c} 0$ with $\tilde{\nu}^0$ as above, then (A1$^+$) holds with $\tilde{f}_0 = f_0$.

Remark 6.6. At least when $r = 2$, the master theorems in Section 2 can be extended to abstract recursions for which the non-degeneracy condition (A4) does not hold and the limiting covariance matrices need not be positive definite. These degenerate cases can be handled by first perturbing the Lipschitz functions f_k and then applying a continuity argument that has some similarities with the proof of Theorem 3.1 in Section 6.8. One of the intermediate steps relies on the fact that $\|W\|_{2\to 2} := \sup_{u \neq 0} \|Wu\|_2/\|u\|_2 = O_c(1)$ for $W \sim \mathrm{GOE}(n)$ as $n \to \infty$ (e.g. Anderson et al., 2010; Knowles and Yin, 2013); see Javanmard and Montanari (2013, Section 4.2.1), Berthier et al. (2020, Section 5.4) and Fan (2022, Appendix D) for further details.

These perturbation arguments do not generalise straightforwardly to d_r convergence results for $r \neq 2$ since $\|W\|_{r\to r} := \sup_{u\neq 0} \|Wu\|_r/\|u\|_r$ is not $O_c(1)$ or even $O_p(1)$ for $r \in [1,2) \cup (2,\infty]$. Indeed, given $r \in [1,2)$ and independent $Z \sim N_n(0, I_n)$ and $\zeta \sim N(0, 1/n)$, Lemma 6.14 yields

$$\|W\|_{r\to r} \geq \|We_1\|_r \overset{d}{=} \|n^{-1/2}Z + \zeta e_1\|_r = n^{1/r-1/2}\|Z\|_{n,r} + o_p(1),$$

where $\|Z\|_{n,r} \xrightarrow{c} \mathbb{E}(|Z_1|^r)^{1/r} \in (0,\infty)$. Moreover, $\|W\|_{r\to r} = \|W\|_{r'\to r'}$ whenever $1/r + 1/r' = 1$.

Remark 6.7. (A5) is a non-vacuous albeit very mild condition. For any Lipschitz $f \colon \mathbb{R}^2 \to \mathbb{R}$, the partial derivative $\frac{\partial f}{\partial x}$ is bounded on its domain of definition, which is a Borel set of full Lebesgue measure. Nevertheless, there are examples of Lipschitz $f \colon \mathbb{R}^2 \to \mathbb{R}$ for which $\frac{\partial f}{\partial x}$ cannot be extended to a function on $\mathbb{R}^2$ that is continuous $(\lambda \otimes \pi)$-almost everywhere (see Remark 7.15). That said, it is inconceivable that such pathological choices of f_k would be made in any practical AMP procedure, where the functions f_k' usually have the property that $\{x \in \mathbb{R} : (x,y) \in D_k\}$ is finite for every $y \in \mathbb{R}$, and hence satisfy (A5).

6.2 Conditional distributions for symmetric AMP

In this subsection, we fix $n \in \mathbb{N}$, and in most places, we suppress the dependence on n of all quantities such as $W \equiv W(n)$ and $h^k \equiv h^k(n)$. When we refer to *orthonormal* sets, it is implicit that the constituent vectors have unit Euclidean norm, i.e. that the underlying inner product is $\langle\cdot\,,\cdot\rangle$, not $\langle\cdot\,,\cdot\rangle_n$. All statements concerning conditional distributions can be understood formally in terms of the rigorous definition of regular conditional probability, as outlined in Section 7.2. The proofs of the results below are given in Section 6.3.

In the setting of Section 2.1, define the $n \times k$ matrices

$$H_k \equiv H_k(n) := \big(h^1 \,\cdots\, h^k\big), \quad M_k \equiv M_k(n) := \big(m^0\; m^1 \,\cdots\, m^{k-1}\big), \quad Y_k \equiv Y_k(n) := \big(y^0\; y^1 \,\cdots\, y^{k-1}\big),$$

where $y^j \equiv y^j(n) := Wm^j = h^{j+1} + b_j m^{j-1}$ for $j = 0, 1, \ldots, k-1$. For convenience, we also define $M_0(n) = Y_0(n) := 0 \in \mathbb{R}^n$. Then the symmetric AMP recursion (2.1) can be rewritten as $WM_k = Y_k$ for $k \in \mathbb{N}$.

For each $0 \leq k \leq n-1$, let $P_k := M_k M_k^+ = M_k(M_k^\top M_k)^+ M_k^\top$ and $P_k^\perp := I_n - P_k$ be the $n \times n$ matrices representing the orthogonal projections onto $\mathrm{Im}(M_k) := \mathrm{span}\{m^j : 0 \leq j \leq k-1\}$ and $V_k := \mathrm{Im}(M_k)^\perp$ respectively, and define $r_k := \mathrm{rank}(M_k) = \dim \mathrm{Im}(M_k)$. Let $\overset{\perp}{m}{}^k := P_k^\perp m^k$ for $0 \leq k \leq n-1$, so that the span of $\overset{\perp}{m}{}^k$ is the orthogonal complement of V_{k+1} within V_k. Furthermore, define $\mathcal{S}_{-1} := \{\emptyset, \Omega\}$ to be the trivial σ-algebra, and for $k \in \mathbb{N}_0$, let

$$\mathcal{S}_k := \sigma(\gamma, m^0, h^j : 1 \leq j \leq k).$$

Then since b_k, m^k are measurable functions of h^k and γ, we see from (2.1) that

$$\mathcal{S}_k = \sigma(\gamma, m^0, y^j : 0 \leq j \leq k-1), \tag{6.3}$$

and that $m^0, \dots, m^k$ and $r_0, \dots, r_{k+1}$ are $\mathcal{S}_k$-measurable for each $-1 \leq k \leq n-1$. (It is not true in general that $\mathbb{P}(r_k = k) = 1$ for all $1 \leq k \leq n-1$, even in recursions (2.1) with non-pathological f_k.)

Our first task is to establish an important fact (Proposition 6.8) that will be used to derive the (regular) conditional distributions of W and h^{k+1} given $\mathcal{S}_k$ in Proposition 6.11 below, for each fixed $k \in \{0, 1, \dots, n-1\}$. We will use the symbol '$\overset{d}{=}|_{\mathcal{S}_k}$' to indicate (almost-sure) equality of conditional distributions given $\mathcal{S}_k$, a notion that is defined formally in Section 7.2.

Proposition 6.8. *Fix $0 \leq k \leq n-1$ and suppose as in (A0) that $W \sim \mathrm{GOE}(n)$ is independent of (m^0, γ). If $\tilde{U}_k$ is any $\mathcal{S}_{k-1}$-measurable $n \times (n - r_k)$ matrix whose columns form an orthonormal basis of V_k, then given $\mathcal{S}_{k-1}$, the matrix $\tilde{U}_k^\top W \tilde{U}_k$ has conditional distribution $\mathrm{GOE}(n - r_k)$ and is conditionally independent of $\mathcal{S}_k$. Consequently, $\tilde{U}_k^\top W \tilde{U}_k$ has conditional distribution $\mathrm{GOE}(n - r_k)$ given $\mathcal{S}_k$, and if $\tilde{W} \sim \mathrm{GOE}(n)$ is independent of $\mathcal{S}_k$, then $\tilde{U}_k^\top W \tilde{U}_k \overset{d}{=}|_{\mathcal{S}_k} \tilde{U}_k^\top \tilde{W} \tilde{U}_k$.*

Remark 6.9. Consider the important special case where $\mathbb{P}(r_k = k) = 1$. Then under the hypotheses of the proposition, $\tilde{U}_k^\top W \tilde{U}_k \sim \mathrm{GOE}(n-k)$ is conditionally independent of $\mathcal{S}_k$ given $\mathcal{S}_{k-1}$, and is independent of $\mathcal{S}_k$.

Remark 6.10. To explicitly construct a (random) $\tilde{U}_k$ with the above properties, consider applying the Gram–Schmidt procedure to $m^0, \dots, m^{k-1}, e_1, \dots, e_n \in \mathbb{R}^n$ (in that order) and retaining only the non-zero vectors in the output (which are all normalised to have unit Euclidean length). This yields an $\mathcal{S}_{k-1}$-measurable orthonormal basis $\tilde{m}^1, \dots, \tilde{m}^n$ of $\mathbb{R}^n$, where $\tilde{m}^1, \dots, \tilde{m}^{r_k}$ are obtained from $m^0, \dots, m^{k-1}$ and therefore span $\mathrm{Im}(M_k)$, while $\tilde{m}^{r_k+1}, \dots, \tilde{m}^n$ span $V_k = \mathrm{Im}(M_k)^\perp$. Thus, we can take $\tilde{U}_k = \bigl(\tilde{m}^{r_k+1} \;\cdots\; \tilde{m}^n\bigr)$.

The main result of this subsection is Proposition 6.11 below, which plays a crucial role in the inductive proof of the AMP master theorems given in Sections 6.4 and 6.5. For each $k \in \{1, \dots, n-1\}$, let

$$\alpha^k \equiv \alpha^k(n) \equiv (\alpha_1^k, \dots, \alpha_k^k) := M_k^+ m^k = (M_k^\top M_k)^+ M_k^\top m^k \in \mathbb{R}^k \tag{6.4}$$

be a vector of projection coefficients satisfying $P_k m^k = M_k \alpha^k = \sum_{\ell=1}^k \alpha_\ell^k m^{\ell-1}$. When M_k has full rank (i.e. when $r_k = k$), note that $\alpha^k = (M_k^\top M_k)^{-1} M_k^\top m^k$ is the unique vector with this property. In addition, let $B_1 := (0, 0) \in \mathbb{R}^2$ and $B_k := \mathrm{diag}(b_0, \dots, b_{k-1}) \in \mathbb{R}^{k \times k}$ for $k \in \{2, \dots, n\}$, so that $Y_k = H_k + (0\ M_{k-1}) B_k$ for all $k \in \{1, \dots, n\}$.

Proposition 6.11. *For $n \in \mathbb{N}$, consider a symmetric AMP recursion (2.1) for which (A0) holds. For $k \in \{0, 1, \dots, n-1\}$, let both $\tilde{W}^k \equiv \tilde{W}^k(n) \sim \mathrm{GOE}(n)$ and $(\tilde{Z}^{k+1}, \tilde{\zeta}^{k+1}) \equiv \bigl(\tilde{Z}^{k+1}(n), \tilde{\zeta}^{k+1}(n)\bigr) \sim N_n(0, I_n) \otimes N(0, 1/n)$ be independent of $\mathcal{S}_k$. Then*

$$W \overset{d}{=}|_{\mathcal{S}_0} \tilde{W}^0 \quad \textit{and} \quad h^1 \overset{d}{=}|_{\mathcal{S}_0} \|m^0\|_n \tilde{Z}^1 + \tilde{\zeta}^1 m^0 =: h^{1,0}, \tag{6.5}$$

and for each $k \in \{1, \dots, n-1\}$, we have

$$W \overset{d}{=}|_{\mathcal{S}_k} W P_k + (W P_k)^\top P_k^\perp + P_k^\perp \tilde{W}^k P_k^\perp = Y_k M_k^+ + (Y_k M_k^+)^\top P_k^\perp + P_k^\perp \tilde{W}^k P_k^\perp \tag{6.6}$$

$$\begin{aligned} h^{k+1} &\overset{d}{=}|_{\mathcal{S}_k} H_k \alpha^k + P_k^\perp(\tilde{W}^k \overset{\perp}{m}{}^k) + \bigl\{(M_k^+)^\top H_k^\top \overset{\perp}{m}{}^k - b_k m^{k-1} + (0\ M_{k-1}) B_k \alpha^k\bigr\} \\ &\overset{d}{=}|_{\mathcal{S}_k} \sum_{\ell=1}^k \alpha_\ell^k h^\ell + \|\overset{\perp}{m}{}^k\|_n (P_k^\perp \tilde{Z}^{k+1}) + \tilde{\zeta}^{k+1} \overset{\perp}{m}{}^k + M_k(M_k^\top M_k)^+ \biggl(v^{k,k} - \sum_{\ell=1}^k \alpha_\ell^k v^{k,\ell-1}\biggr) \qquad (6.7) \\ &=: h^{k+1,k}, \end{aligned}$$

where $v^{k,\ell} \equiv v^{k,\ell}(n) := H_k^\top m^\ell - b_\ell \, M_k^\top m^{\ell-1} \in \mathbb{R}^k$ *for* $\ell \in \{0, \dots, k\}$.

The crux of the proof of Proposition 6.11 is to establish (6.6), which characterises the conditional distribution of W given $\mathcal{S}_k$. It is intuitively helpful to think of this as being obtained by conditioning W on the 'linear constraints' $Wm^0 = y^0, \dots, Wm^{k-1} = y^{k-1}$. However, since $m^1, \dots, m^{k-1}$ are random and depend on W, this heuristic argument is not sufficient on its own to constitute a formal proof of Proposition 6.11. For the benefit of readers interested in the technicalities, we give a more detailed explanation below.

Observe that for fixed $k \in \mathbb{N}$ and deterministic $y, a^0, a^1, \dots, a^k \in \mathbb{R}^n$, the event $\Omega_{y,a^0,\dots,a^k} := \{\gamma = y, m^0 = a^0, h^1 = a^1, \dots, h^k = a^k\}$ can be expressed as

$$\Omega_{y,a^0,\dots,a^k} = \{\gamma = y, m^0 = a^0, Wt^j = z^j \text{ for all } 0 \le j \le k-1\} = \{\gamma = y, m^0 = a^0, WT_k = Z_k\}, \quad (6.8)$$

where $t^j := f_j(a^j, y) \in \mathbb{R}^n$ and $z^j := a^{j+1} + \langle f_j'(a^j, y)\rangle_n \, f_{j-1}(a^{j-1}, y)$ for $0 \le j \le k-1$, and $T_k := (t^0 \; t^1 \; \cdots \; t^{k-1})$ and $Z_k := (z^0 \; z^1 \; \cdots \; z^{k-1})$ are fixed $n \times k$ matrices. Now for $W \sim \mathrm{GOE}(n)$ and any fixed $T \in \mathbb{R}^{n \times k}$ of rank p, we can derive the conditional distribution of W given WT by writing

$$W = WP + \big(P + P^\perp\big)^\top WP^\perp = WP + (WP)^\top P^\perp + P^\perp W P^\perp, \quad (6.9)$$

where $P := TT^+$ and $P^\perp := I_n - TT^+$ represent the orthogonal projections onto $\mathrm{Im}(T)$ and $\mathrm{Im}(T)^\perp$ respectively. The first two terms on the right hand side of (6.9) are measurable functions of $WP = (WT)T^+$ (and hence WT), while the third term $P^\perp W P^\perp$ is independent of WT. Thus, $\mathbb{E}(W \,|\, WT) = WP + (WP)^\top P^\perp$. Moreover, we can write $P^\perp = \tilde{U}\tilde{U}^\top$, where the columns of $\tilde{U}$ form an orthonormal basis for $\mathrm{Im}(T)^\perp$, so that $P^\perp W P^\perp = \tilde{U}(\tilde{U}^\top W \tilde{U})\tilde{U}^\top$, and $\tilde{U}^\top W \tilde{U} \sim \mathrm{GOE}(n-p)$ is independent of WT. For $Z \in \mathbb{R}^{n \times k}$, this enables us to interpret 'the conditional distribution of W given $WT = Z$' as the distribution of

$$ZT^+ + (ZT^+)^\top P^\perp + \tilde{U}^\top \tilde{W} \tilde{U},$$

where $\tilde{W} \sim \mathrm{GOE}(n-p)$. We denote this distribution by $\mathcal{L}_Z(T)$.

In view of (6.8) and the assumption that W is independent of (m^0, γ) in (A0), it is then tempting to argue heuristically that

$$\begin{aligned} W \,|\, \text{`}\{\gamma = y, m^0 = a^0, h^1 = a^1, \dots, h^k = a^k\}\text{'} &\overset{d}{=} W \,|\, \text{`}\{\gamma = y, m^0 = a^0, WT_k = Z_k\}\text{'} \\ &\overset{d}{=} W \,|\, \text{`}\{WT_k = Z_k\}\text{'} \sim \mathcal{L}_{Z_k}(T_k), \end{aligned}$$

and conclude on this basis that W has (regular) conditional distribution $\mathcal{L}_\omega \equiv \mathcal{L}_{Y_k(\omega)}\big(M_k(\omega)\big)$ given $\mathcal{S}_k = \sigma(\gamma, m^0, h^1, \dots, h^k)$, noting that $M_k(\omega) = T_k$ and $Y_k(\omega) = Z_k$ for $\omega \in \Omega_{y,a^0,\dots,a^k}$. However, this line of reasoning appears to involve conditioning explicitly on an event of potentially zero probability, and is not formally justified by the above argument; cf. the Borel paradox (Dudley, 2002, pp. 350–351) for the associated hazards.

As mentioned above, the issue is that M_k is random and is in general not independent of W, whereas the distributional claims in the previous paragraph relied on the fact that T was fixed. Nevertheless, the key point is that the randomness of M_k and its dependence on W turn out not to cause irreconcilable difficulties, due to the conditional independence established in Proposition 6.8. It follows from this result that $\mathbb{E}(W \,|\, \mathcal{S}_k) = WP_k + (WP_k)^\top P_k^\perp$, so the conditional distributional equality (6.6) in Proposition 6.11 and the decomposition (6.11) in its proof are the appropriate analogues of (6.9).

6.3 Proofs of results in Section 6.2

A key ingredient in the proof of Proposition 6.8 is Lemma 6.12 below, which extends the orthogonal invariance property of the GOE(n) distribution. Given a finite collection of disjoint measurable spaces $(\mathcal{X}_1, \mathcal{A}_1), \dots, (\mathcal{X}_m, \mathcal{A}_m)$, we equip the disjoint union $\bigsqcup_{k=1}^m \mathcal{X}_k$ with the σ-algebra $\big\{\bigsqcup_{k=1}^m A_k : A_k \in \mathcal{A}_k \text{ for all } k\big\}$.

Lemma 6.12. *Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra and let $X\colon (\Omega, \mathcal{F}, \mathbb{P}) \to \bigsqcup_{k=1}^{n} \mathbb{R}^{k\times k}$ be a measurable function. Suppose that there is a partition of Ω into disjoint events $\Omega_1, \ldots, \Omega_m \in \mathcal{G}$ such that for each $k = 1, \ldots, m$, the map X takes values in $\mathbb{R}^{n_k \times n_k}$ on Ω_k and has conditional distribution $\mathrm{GOE}(n_k)$ given $\mathcal{G}$ on Ω_k, for some (deterministic) $n_k \in \{1, \ldots, n\}$. Moreover, let $Q = (Q_1\ Q_2)\colon (\Omega, \mathcal{F}, \mathbb{P}) \to \bigsqcup_{k=1}^{n} \mathbb{O}_k$ be a $\mathcal{G}$-measurable function such that on each event Ω_k, the map Q takes values in $\mathbb{O}_{n_k}$, and Q_1, Q_2 have ℓ_k and $n_k - \ell_k$ columns respectively, for some (deterministic) $\ell_k \in \{1, \ldots, n_k - 1\}$. Then, given $\mathcal{G}$, we have the following:*

(a) $Q^\top XQ$ has conditional distribution $\mathrm{GOE}(n_k)$ on Ω_k for every $k = 1, \ldots, m$;

(b) $Q_2^\top XQ_2$ has conditional distribution $\mathrm{GOE}(n_k - \ell_k)$ on Ω_k for every $k = 1, \ldots, m$;

(c) $Q^\top XQ_1$ and $Q_2^\top XQ_2$ are conditionally independent.

Remark 6.13. Note that if $n_1 = \cdots = n_m = n$, then under the first condition of the lemma, it follows from Remark 7.4 that X has unconditional distribution $\mathrm{GOE}(n)$ and is independent of $\mathcal{G}$. Thus, in the instructive special case where $m = 1$ and $1 \le \ell_1 < n = n_1$, the result above simplifies to the following: suppose that $X \sim \mathrm{GOE}(n)$, and is independent of $\mathcal{G}$, and moreover that $Q = (Q_1\ Q_2)\colon (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{O}_n$ is a $\mathcal{G}$-measurable map such that Q_1, Q_2 have ℓ_1 and $n - \ell_1$ columns respectively. Then

(a) $Q^\top XQ \sim \mathrm{GOE}(n)$ and is independent of $\mathcal{G}$;

(b) $Q_2^\top XQ_2 \sim \mathrm{GOE}(n - \ell_1)$ and is independent of $\mathcal{G}$;

(c) $Q^\top XQ_1$ and $Q_2^\top XQ_2$ are independent, and also conditionally independent given $\mathcal{G}$.

Proof of Lemma 6.12. (a) For $\ell = 1, \ldots, n$, let $\mathcal{A}_\ell$ and $\mathcal{B}_\ell$ be the Borel σ-algebras on $\mathcal{X}_\ell := \mathbb{R}^{\ell\times\ell}$ and $\mathcal{Y}_\ell := \mathbb{O}_\ell$ respectively. Define $\phi_\ell\colon \mathcal{X}_\ell \times \mathcal{Y}_\ell \to \mathcal{X}_\ell$ by $\phi_\ell(M, J) := J^\top MJ$, and for $J \in \mathbb{O}_\ell$, let $\iota_J\colon \mathcal{X}_\ell \to \mathcal{X}_\ell \times \mathcal{Y}_\ell$ be the map $M \mapsto (M, J)$. Then the orthogonal invariance property of $\mathrm{GOE}(\ell)$ can be restated as $\mathrm{GOE}(\ell) = \mathrm{GOE}(\ell) \circ (\phi_\ell \circ \iota_J)^{-1}$ for every $J \in \mathbb{O}_\ell$. Thus, observing that $\phi_{n_k}(X, Q) = Q^\top XQ$ on Ω_k, and applying Lemma 7.6(b) to ϕ_{n_k}, we see that $Q^\top XQ$ has conditional distribution $\mathrm{GOE}(n_k)$ given $\mathcal{G}$ on Ω_k, as required.

(b) For $k = 1, \ldots, m$, let $\psi_k\colon \mathcal{X}_{n_k} \to \mathcal{X}_{n_k - \ell_k}$ denote the map that extracts the lower-right $(n_k - \ell_k) \times (n_k - \ell_k)$ block of entries of an $n_k \times n_k$ matrix. Then $\psi_k(W) \sim \mathrm{GOE}(n_k - \ell_k)$ whenever $W \sim \mathrm{GOE}(n_k)$, so $\mathrm{GOE}(n_k - \ell_k) = \mathrm{GOE}(n_k) \circ \psi_k^{-1} = \mathrm{GOE}(n_k) \circ (\phi_{n_k} \circ \iota_J)^{-1} \circ \psi_k^{-1}$ for every $J \in \mathbb{O}_{n_k}$. We can therefore apply Lemma 7.6(b) to $\psi_k \circ \phi_{n_k}$ to conclude that $Q_2^\top XQ_2$ has conditional distribution $\mathrm{GOE}(n_k - \ell_k)$ given $\mathcal{G}$ on Ω_k.

(c) For $\omega \in \Omega$, let P_ω, Q_ω and R_ω respectively denote the conditional distributions of $Q^\top XQ_1$, $Q_2^\top XQ_2$ and $(Q^\top XQ_1, Q_2^\top XQ_2)$ given $\mathcal{G}$. For $k = 1, \ldots, m$, let $\tilde{\psi}_k\colon \mathcal{X}_{n_k} \to \mathbb{R}^{n_k \times \ell_k}$ denote the map that extracts the first ℓ_k columns of a $n_k \times n_k$ matrix. Now define $\Psi_k\colon \mathcal{X}_{n_k} \to \mathbb{R}^{n_k\times\ell_k} \times \mathcal{X}_{n_k - \ell_k}$ by $\Psi_k(M) := \bigl(\tilde{\psi}_k(M), \psi_k(M)\bigr)$. Then $\tilde{\psi}_k(W)$ and $\psi_k(W)$ are independent whenever $W \sim \mathrm{GOE}(n_k)$, so $\mathrm{GOE}(n_k) \circ \Psi_k^{-1} = \bigl(\mathrm{GOE}(n_k) \circ \tilde{\psi}_k^{-1}\bigr) \otimes \bigl(\mathrm{GOE}(n_k) \circ \psi_k^{-1}\bigr)$. Since $(Q^\top XQ_1, Q_2^\top XQ_2) = (\Psi_k \circ \phi_k)(X, Q)$ on Ω_k, we may apply Lemma 7.6(b) to $\tilde{\psi}_k \circ \phi_{n_k}$, $\psi_k \circ \phi_{n_k}$ and $\Psi_k \circ \phi_{n_k}$ to deduce that $R_\omega = P_\omega \otimes Q_\omega$ for all $\omega \in \Omega_k$. Since $k \in \{1, \ldots, m\}$ was arbitrary, we conclude that $R_\omega = P_\omega \otimes Q_\omega$ for all $\omega \in \Omega = \bigsqcup_{k=1}^{m} \Omega_k$, which together with Lemma 7.9(b) implies that $Q^\top XQ_1$ and $Q_2^\top XQ_2$ are conditionally independent given $\mathcal{G}$. □

Proof of Proposition 6.8. We argue by induction on $k \in \{0, 1, \ldots, n-1\}$. The case $k = 0$ is trivial since $W \sim \mathrm{GOE}(n)$ and is independent of (m^0, γ) by assumption. For a general $1 \le k \le n-1$ (when $n \ge 2$), let $\tilde{U}_{k-1}$ be any $\mathcal{S}_{k-2}$-measurable $n \times (n - r_{k-1})$ matrix whose columns form an orthonormal basis of V_{k-1}, and fix an arbitrary $\mathcal{S}_{k-1}$-measurable $n \times (n - r_k)$ matrix $\tilde{U}_k$ whose columns form an orthonormal basis of V_k. Moreover, let $E \equiv E_{k-1}$ be the event $\{\overset{\perp}{m}{}^{k-1} \neq 0\} = \{r_k = r_{k-1} + 1\} \in \mathcal{S}_{k-1}$, and note that $n - r_{k-1} \ge n - k + 1 \ge 2$.

Next, define an $\mathcal{S}_{k-1}$-measurable $n \times (n - r_{k-1})$ matrix $\check{U}$ by setting $\check{U} := (\overset{\perp}{m}^{k-1} \ \tilde{U}_k)$ on E and $\check{U} := \tilde{U}_k$ on $E^c = \{\overset{\perp}{m}^{k-1} = 0\}$. Letting $\breve{U}$ be the $\mathcal{S}_{k-1}$-measurable $n \times (n - r_{k-1} - 1)$ matrix obtained by removing the first column $\check{U}e_1$ of $\check{U}$, we therefore have $\tilde{U}_k = \breve{U}$ on E and $\tilde{U}_k = \check{U}$ on E^c. Now $\check{U}, \tilde{U}_{k-1}$ have orthonormal columns that span V_{k-1}, so $Q := \tilde{U}_{k-1}^\top \check{U}$ is an $\mathcal{S}_{k-1}$-measurable orthogonal $(n - r_{k-1}) \times (n - r_{k-1})$ matrix such that $\check{U}^\top W \check{U} = Q^\top \big(\tilde{U}_{k-1}^\top W \tilde{U}_{k-1}\big) Q$. By the inductive hypothesis, $\tilde{U}_{k-1}^\top W \tilde{U}_{k-1}$ has conditional distribution $\mathrm{GOE}(n - r_{k-1})$ given $\mathcal{S}_{k-1}$, so it follows from parts (a) and (b) respectively of Lemma 6.12 (with $\ell \equiv 1$ and $N = n - r_{k-1} \geq 2$) that $\check{U}^\top W \check{U}$ and $\breve{U}^\top W \breve{U}$ have conditional distributions $\mathrm{GOE}(n - r_{k-1})$ and $\mathrm{GOE}(n - r_{k-1} - 1)$ respectively given $\mathcal{S}_{k-1}$. Since $\tilde{U}_k^\top W \tilde{U}_k = \check{U}^\top W \check{U}$ on $E^c = \{r_k = r_{k-1}\} \in \mathcal{S}_{k-1}$ and $\tilde{U}_k^\top W \tilde{U}_k = \breve{U}^\top W \breve{U}$ on $E = \{r_k = r_{k-1} + 1\} \in \mathcal{S}_{k-1}$, we deduce from Lemma 7.6(a) that $\tilde{U}_k^\top W \tilde{U}_k$ has conditional distribution $\mathrm{GOE}(n - r_k)$ given $\mathcal{S}_{k-1}$, as required.

In addition, it holds trivially that 0 and $\check{U}^\top W \check{U}$ are conditionally independent given $\mathcal{S}_{k-1}$, and Lemma 6.12(c) implies that $\check{U}^\top W (\check{U}e_1) = Q^\top (\tilde{U}_{k-1}^\top W \tilde{U}_{k-1}) Q e_1$ and $\breve{U}^\top W \breve{U}$ are also conditionally independent given $\mathcal{S}_{k-1}$. Since $W \overset{\perp}{m}^{k-1} = 0$ on E^c and $\check{U}e_1 = \overset{\perp}{m}^{k-1}$ on E, an application of Lemma 7.9(a) shows that $\check{U}^\top W \overset{\perp}{m}^{k-1}$ (and hence $\sigma(\mathcal{S}_{k-1}, \check{U}^\top W \overset{\perp}{m}^{k-1})$ by Lemma 7.8) is conditionally independent of $\tilde{U}_k^\top W \tilde{U}_k$ given $\mathcal{S}_{k-1}$. Moreover, $WP_{k-1} = WM_{k-1}M_{k-1}^+ = Y_{k-1}M_{k-1}^+$, m^{k-1}, $\overset{\perp}{m}^{k-1}$, $\check{U}$ and $P_{k-1}^\perp = \check{U}\check{U}^\top$ are $\mathcal{S}_{k-1}$-measurable, so

$$\begin{aligned} y^{k-1} = W\big(P_{k-1} + P_{k-1}^\perp\big) m^{k-1} &= (WP_{k-1}) m^{k-1} + \big(P_{k-1} + P_{k-1}^\perp\big)^\top W \overset{\perp}{m}^{k-1} \\ &= (WP_{k-1}) m^{k-1} + (WP_{k-1})^\top \overset{\perp}{m}^{k-1} + \check{U}\big(\check{U}^\top W \overset{\perp}{m}^{k-1}\big) \end{aligned} \tag{6.10}$$

is measurable with respect to $\sigma(\mathcal{S}_{k-1}, \check{U}^\top W \overset{\perp}{m}^{k-1})$. Thus, given $\mathcal{S}_{k-1}$, we conclude that $\tilde{U}_k^\top W \tilde{U}_k$ is conditionally independent of y^{k-1}, and hence conditionally independent of $\mathcal{S}_k = \sigma(\mathcal{S}_{k-1}, y^{k-1})$ by Lemma 7.8 and (6.3). Therefore, since $\tilde{U}_k^\top W \tilde{U}_k$ has conditional distribution $\mathrm{GOE}(n - r_k)$ given $\mathcal{S}_{k-1}$, it also has conditional distribution $\mathrm{GOE}(n - r_k)$ given $\sigma(\mathcal{S}_{k-1}, \mathcal{S}_k) = \mathcal{S}_k$.

Finally, it remains to show that if $\tilde{W} \sim \mathrm{GOE}(n)$ is independent of $\mathcal{S}_k$, then $\tilde{U}_k^\top \tilde{W} \tilde{U}_k$ also has conditional distribution $\mathrm{GOE}(n - r_k)$ given $\mathcal{S}_k$. To see this, let $\tilde{m}^1, \dots, \tilde{m}^{r_k}$ be an $\mathcal{S}_k$-measurable orthonormal basis of $\mathrm{Im}(M_k)$, obtained for example by applying the Gram–Schmidt procedure to $m^0, \dots, m^{k-1}$, as in Remark 6.10 above. Then taking $Q_1 = \big(\tilde{m}^1 \ \cdots \ \tilde{m}^{r_k}\big)$ and $Q_2 = \tilde{U}_k$, we see that $Q = (Q_1 \ Q_2)$ satisfies the hypotheses of Lemma 6.12 with $\Omega_j = \{r_k = j\} \in \mathcal{S}_k$ and $\ell_j = j \leq n = n_j$ for $j = 1, \dots, k$. The desired conclusion now follows directly from Lemma 6.12(b), and this completes the inductive step. □

As mentioned above, the proof of Proposition 6.11 relies crucially on the final assertion in Proposition 6.8. To obtain the conditional distributional equalities in (6.5) and (6.7), we will also apply the following elementary fact.

Lemma 6.14. *If $W \sim \mathrm{GOE}(n)$ and $u \in \mathbb{R}^n$ is fixed, then $Wu \overset{d}{=} \|u\|_n Z + \zeta u$, where $Z \sim N_n(0, I_n)$ and $\zeta \sim N(0, 1/n)$ are independent.*

Proof of Lemma 6.14. The result holds trivially when $u = 0$, and is also true when $u = e_1$ since $We_1 \sim N_n\big(0, \mathrm{diag}(2/n, 1/n, \dots, 1/n)\big)$. For a general $u \in \mathbb{R}^n \setminus \{0\}$, let $Q \in \mathbb{R}^{n \times n}$ be an orthogonal matrix with $Qe_1 = u/\|u\|$, so that $Q^\top u = \|u\| e_1$. Then

$$Wu \overset{d}{=} QWQ^\top u = Q(We_1)\|u\| \overset{d}{=} Q(\|e_1\|_n Z + \zeta e_1)\|u\| \overset{d}{=} \|u\|_n Z + \zeta u,$$

as required, where we have used the orthogonal invariance of $W \sim \mathrm{GOE}(n)$, the result for e_1 and the orthogonal invariance of $Z \sim N_n(0, I_n)$ respectively to obtain the distributional equalities above. □

Proof of Proposition 6.11. We start by proving (6.6) for every $k \in \{0, 1, \dots, n-1\}$. Let $\tilde{U}_k$ be any $\mathcal{S}_{k-1}$-measurable $n \times (n - r_k)$ matrix whose columns form an orthonormal basis of V_k; see Remark 6.10

for a specific construction of $\tilde{U}_k$. Similarly to (6.10) in the proof of Proposition 6.8, we can write

$$\begin{aligned} W &= WP_k + \bigl(P_k + P_k^\perp\bigr)^\top W P_k^\perp \\ &= WP_k + (WP_k)^\top P_k^\perp + P_k^\perp W P_k^\perp \\ &= WP_k + (WP_k)^\top P_k^\perp + \tilde{U}_k(\tilde{U}_k^\top W \tilde{U}_k)\tilde{U}_k^\top \\ &\overset{d}{=}|_{\mathcal{S}_k} WP_k + (WP_k)^\top P_k^\perp + \tilde{U}_k(\tilde{U}_k^\top \tilde{W}^k \tilde{U}_k)\tilde{U}_k^\top = WP_k + (WP_k)^\top P_k^\perp + P_k^\perp \tilde{W}^k P_k^\perp, \end{aligned} \tag{6.11}$$

where $\tilde{W}^k \sim \mathrm{GOE}(n)$ is independent of $\mathcal{S}_k$. To justify the key distributional equality after (6.11), we can apply Lemma 7.6(c); indeed, note that $WP_k = Y_k M_k^+$, $\tilde{U}_k$ and $P_k^\perp = \tilde{U}_k\tilde{U}_k^\top$ are $\mathcal{S}_k$-measurable, and that $\tilde{U}_k^\top W \tilde{U}_k \overset{d}{=}|_{\mathcal{S}_k} \tilde{U}_k^\top \tilde{W}^k \tilde{U}_k$ by the final assertion of Proposition 6.8. By replacing WP_k with $Y_k M_k^+$ in the display above, we obtain (6.6) for every $k \in \{0, 1, \ldots, n-1\}$, as desired. Since $I - P_0 = P_0^\perp = I_n$, this specialises to $W \overset{d}{=}|_{\mathcal{S}_0} \tilde{W}^0$ when $k = 0$, which is the first part of (6.5).

Using (6.6), we now derive the conditional distribution of h^{k+1} given $\mathcal{S}_k$ for $k \in \{0, 1, \ldots, n-1\}$. When $k = 0$, we have $h^1 = Wm^0$, so the associated identity in (6.5) follows directly from the first part of (6.5), Lemma 6.14 and Lemma 7.6(c). Turning now to (6.7) with $k \geq 1$, we have $h^{k+1} = Wm^k - b_k m^{k-1}$, where b_k, m^{k-1} are $\mathcal{S}_k$-measurable, so we can deduce from (6.6) and Lemma 7.6(c) that

$$\begin{aligned} h^{k+1} &\overset{d}{=}|_{\mathcal{S}_k} Y_k M_k^+ m^k + (Y_k M_k^+)^\top P_k^\perp m^k + P_k^\perp \tilde{W}^k P_k^\perp m^k - b_k m^{k-1} \\ &= Y_k \alpha^k + (Y_k M_k^+)^\top \overset{\perp}{m}{}^k + P_k^\perp(\tilde{W}^k \overset{\perp}{m}{}^k) - b_k m^{k-1} \\ &= H_k \alpha^k + (0\ M_{k-1}) B_k \alpha^k + (H_k M_k^+)^\top \overset{\perp}{m}{}^k + P_k^\perp(\tilde{W}^k \overset{\perp}{m}{}^k) - b_k m^{k-1}. \end{aligned} \tag{6.12}$$

Indeed, to obtain the final equality above, observe that $Y_k = H_k + (0\ M_{k-1}) B_k$ and $B_k^\top (0\ M_{k-1})^\top \overset{\perp}{m}{}^k = B_k^\top (0\ M_{k-1})^\top P_k^\perp m^k = 0$ in view of the fact that $P_k^\perp M_{k-1} = 0$. Since $\overset{\perp}{m}{}^k$ is $\mathcal{S}_k$-measurable and $\tilde{W}^k \sim \mathrm{GOE}(n)$ is independent of $\mathcal{S}_k$ (and therefore has conditional distribution $\mathrm{GOE}(n)$ given $\mathcal{S}_k$), it follows from Lemmas 6.14 and 7.6(b) that $\tilde{W}^k \overset{\perp}{m}{}^k \overset{d}{=}|_{\mathcal{S}_k} \|\overset{\perp}{m}{}^k\|_n \tilde{Z}^{k+1} + \tilde{\zeta}^{k+1} \overset{\perp}{m}{}^k$. Now since $P_k^\perp$ and all the other summands in (6.12) are $\mathcal{S}_k$-measurable, a further application of Lemma 7.6(c) shows that the random variable in (6.12) and

$$\begin{aligned} &H_k \alpha^k + P_k^\perp\bigl(\|\overset{\perp}{m}{}^k\|_n \tilde{Z}^{k+1} + \tilde{\zeta}^{k+1} \overset{\perp}{m}{}^k\bigr) + (H_k M_k^+)^\top \overset{\perp}{m}{}^k - \bigl\{b_k m^{k-1} - (0\ M_{k-1}) B_k \alpha^k\bigr\} \\ &= \sum_{\ell=1}^k \alpha_\ell^k\, h^\ell + P_k^\perp\bigl(\|\overset{\perp}{m}{}^k\|_n \tilde{Z}^{k+1} + \tilde{\zeta}^{k+1} \overset{\perp}{m}{}^k\bigr) + (H_k M_k^+)^\top \overset{\perp}{m}{}^k - \Biggl(b_k m^{k-1} - \sum_{\ell=1}^k \alpha_\ell^k\, b_{\ell-1} m^{\ell-2}\Biggr) \end{aligned} \tag{6.13}$$

are identically distributed given $\mathcal{S}_k$. Finally, recall that $\overset{\perp}{m}{}^k = (I - P_k)\, m^k = m^k - \sum_{\ell=1}^k \alpha_\ell^k\, m^{\ell-1}$, and that $(M_k^+)^\top M_k^\top m^\ell = P_k^\top m^\ell = P_k m^\ell = m^\ell$ for all $0 \leq \ell \leq k-1$ by the definition of the projection matrix $P_k = M_k M_k^+$. It follows that $P_k^\perp(\tilde{\zeta}^{k+1} \overset{\perp}{m}{}^k) = \tilde{\zeta}^{k+1} \overset{\perp}{m}{}^k$ and

$$(M_k^+)^\top H_k^\top \overset{\perp}{m}{}^k = (M_k^+)^\top \Biggl(H_k^\top m^k - \sum_{\ell=1}^k \alpha_\ell^k\, H_k^\top m^{\ell-1}\Biggr)$$

$$b_k m^{k-1} - \sum_{\ell=1}^k b_{\ell-1} \alpha_\ell^k\, m^{\ell-2} = (M_k^+)^\top \Biggl(b_k\, M_k^\top m^{k-1} - \sum_{\ell=1}^k \alpha_\ell^k\, b_{\ell-1}\, M_k^\top m^{\ell-2}\Biggr).$$

Thus, since $(M_k^+)^\top = M_k(M_k^\top M_k)^+$, the random variable $h^{k+1,k}$ defined in (6.7) is identical to that in (6.13), so we conclude from (6.12) that $h^{k+1} \overset{d}{=}|_{\mathcal{S}_k} h^{k+1,k}$, as required. □

6.4 Proof outline for the AMP master theorems in Section 2.1

Recalling the definition (2.6) of the limiting covariance matrices $\bar{\mathrm{T}}^{[k]} \in \mathbb{R}^{k \times k}$ in Theorem 2.3, we first outline a standard construction of a single random sequence $(\bar{G}_k : k \in \mathbb{N})$ satisfying $(\bar{G}_1, \ldots, \bar{G}_k) \sim N_k(0, \bar{\mathrm{T}}^{[k]})$ for each k. Let $\bar{\mathrm{T}}^{[k],k+1} := (\bar{\mathrm{T}}_{1,k+1}, \ldots, \bar{\mathrm{T}}_{k,k+1}) \in \mathbb{R}^k$ and

$$\bar{\alpha}^k \equiv (\bar{\alpha}_1^k, \ldots, \bar{\alpha}_k^k) := \bigl(\bar{\mathrm{T}}^{[k]}\bigr)^{-1}\, \bar{\mathrm{T}}^{[k],k+1} \in \mathbb{R}^k \tag{6.14}$$

for each $k \in \mathbb{N}$, where the latter is well-defined since $\bar{\mathrm{T}}^{[k]}$ is positive definite under (A4) by Lemma 2.2. It is easily verified that if $(G_1, \dots, G_{k+1}) \sim N_{k+1}(0, \bar{\mathrm{T}}^{[k+1]})$, then $G_{[k]} := (G_1, \dots, G_k)$ and $\xi_{k+1} := G_{k+1} - G_{[k]}^\top \bar{\alpha}^k = G_{k+1} - \sum_{\ell=1}^k \bar{\alpha}_\ell^k \, G_\ell$ are uncorrelated and hence independent. This means that

$$G_{[k]}^\top \bar{\alpha}^k = \sum_{\ell=1}^k \bar{\alpha}_\ell^k \, G_\ell = \mathbb{E}(G_{k+1} \,|\, G_1, \dots, G_k)$$

for each k. Moreover, since $\bar{\mathrm{T}}^{k+1} = \mathrm{Cov}(G_1, \dots, G_{k+1})$ is positive definite and ξ_{k+1} is a non-trivial linear combination of $G_1, \dots, G_{k+1}$, it follows under (A4) that

$$\begin{aligned} 0 < \mathrm{Var}(\xi_{k+1}) &= \mathrm{Var}(\xi_{k+1} \,|\, G_1, \dots, G_k) = \mathrm{Var}(G_{k+1} \,|\, G_1, \dots, G_k) \\ &= \mathrm{Var}(G_{k+1}) - \mathrm{Var}\big(G_{[k]}^\top \bar{\alpha}^k\big) \\ &= \bar{\mathrm{T}}_{k+1,k+1} - (\bar{\alpha}^k)^\top \bar{\mathrm{T}}^{[k]} \, \bar{\alpha}^k \\ &= \tau_{k+1}^2 - (\bar{\mathrm{T}}^{[k],k+1})^\top \big(\bar{\mathrm{T}}^{[k]}\big)^{-1} \, \bar{\mathrm{T}}^{[k],k+1} =: \overset{\perp}{\tau}{}_{k+1}^2 \end{aligned} \tag{6.15}$$

for $k \in \mathbb{N}$, so that $\overset{\perp}{\tau}_{k+1} \in (0, \infty)$ satisfies $\overset{\perp}{\tau}{}_{k+1}^2 = \mathrm{Var}(\xi_{k+1}) \leq \mathrm{Var}(G_{k+1}) = \tau_{k+1}^2$. Now let $\bar{G}_1 \sim N(0, \tau_1^2)$, and for $k \in \mathbb{N}$, inductively define

$$\bar{G}_{k+1} := \sum_{\ell=1}^k \bar{\alpha}_\ell^k \, \bar{G}_\ell + \overset{\perp}{\tau}_{k+1} \overset{\perp}{\zeta}_{k+1}, \tag{6.16}$$

where $\overset{\perp}{\zeta}_{k+1} \sim N(0,1)$ is independent of $(\bar{G}_1, \dots, \bar{G}_k)$. Then $(\bar{G}_k : k \in \mathbb{N})$ is a random sequence with $(\bar{G}_1, \dots, \bar{G}_k) \sim N_k(0, \bar{\mathrm{T}}^{[k]})$ for each k, as desired. With the above definitions in place, we record here some key identities. In view of (2.6), we certainly have

$$\mathrm{Cov}(\bar{G}_k, \bar{G}_\ell) = \mathbb{E}(\bar{G}_k \bar{G}_\ell) = \bar{\mathrm{T}}_{k,\ell} = \begin{cases} \tau_1^2 & \text{if } k = \ell = 1 \\ \mathbb{E}\big(F_0(\bar{\gamma}) \cdot f_{k-1}(\bar{G}_{k-1}, \bar{\gamma})\big) & \text{if } k > \ell = 1 \\ \mathbb{E}\big(f_{\ell-1}(\bar{G}_{\ell-1}, \bar{\gamma}) \cdot f_{k-1}(\bar{G}_{k-1}, \bar{\gamma})\big) & \text{if } k \geq \ell \geq 2, \end{cases} \tag{6.17}$$

where $f_1, f_2, \dots$ are the Lipschitz functions in the AMP recursion (2.1), and τ_1 and F_0 are as in (A2) and (A3) respectively. This fact underlies an important assertion (Proposition 6.16(e) below) in our inductive proof of the master theorems. Moreover, for $k, \ell \in \mathbb{N}$ and any Lipschitz function $\varphi \colon \mathbb{R} \to \mathbb{R}$ with weak derivative φ', we have

$$\mathbb{E}\big(\bar{G}_k \, \varphi(\bar{G}_\ell)\big) = \mathbb{E}\big(\varphi'(\bar{G}_\ell)\big)\mathbb{E}(\bar{G}_k \bar{G}_\ell) = \mathbb{E}\big(\varphi'(\bar{G}_\ell)\big)\bar{\mathrm{T}}_{k,\ell}. \tag{6.18}$$

This follows from Stein's lemma, a general formulation of which can be found in Tsybakov (2009, Lemma 3.6) and Lemma 6.20.

Lemma 6.15 (Stein's lemma)**.** *If $Z \sim N(0, \sigma^2)$ and $\varphi \colon \mathbb{R} \to \mathbb{R}$ is an absolutely continuous function with weak derivative φ' such that $\varphi'(Z)$ is integrable, then $\mathbb{E}\big(Z\varphi(Z)\big) = \sigma^2 \, \mathbb{E}\big(\varphi'(Z)\big)$.*

Indeed, the first equality in (6.18) follows from Lemma 6.15 upon writing $\bar{G}_k = (\bar{\mathrm{T}}_{k,\ell}/\bar{\mathrm{T}}_{\ell,\ell}) \, \bar{G}_\ell + \xi_{k\ell}$, where $\xi_{k\ell}$ has zero mean and is independent of $\bar{G}_\ell$, so that $\mathbb{E}\big(\xi_{k\ell} \, \varphi(\bar{G}_\ell)\big) = 0$.

Our choice of functions $(f_k)_{k=0}^\infty$ and $(f_k')_{k=0}^\infty$ in (2.1) ensures that for any fixed $y \in \mathbb{R}$, we can take $\varphi = f_\ell(\cdot\,, y)$ and $\varphi' = f_\ell'(\cdot\,, y)$ in (6.18) to see that $\mathbb{E}\big(\bar{G}_k \, f_\ell(\bar{G}_\ell, y)\big) = \mathbb{E}\big(f_\ell'(\bar{G}_\ell, y)\big)\mathbb{E}(\bar{G}_k \bar{G}_\ell)$ for $k, \ell \in \mathbb{N}$. We deduce from this (and Lemma 7.7) that if $\bar{\gamma} \sim \pi$ is independent of $\bar{G}_1, \bar{G}_2, \dots$, then

$$\mathbb{E}\big(\bar{G}_k \, f_\ell(\bar{G}_\ell, \bar{\gamma})\big) = \mathbb{E}\big(f_\ell'(\bar{G}_\ell, \bar{\gamma})\big)\mathbb{E}(\bar{G}_k \bar{G}_\ell) = \bar{b}_\ell \, \bar{\mathrm{T}}_{k,\ell} \tag{6.19}$$

for all $k, \ell \in \mathbb{N}$, where $\bar{b}_\ell := \mathbb{E}\big(f_\ell'(\bar{G}_\ell, \bar{\gamma})\big)$. This forms part of assertion (f) in Proposition 6.16 below.

To complete our technical preparations for the main derivations below, we will set up a more explicit connection between the Gaussian variables $\bar{G}_{k+1} \sim N(0, \tau_{k+1}^2)$ in (6.16) and the random vectors

$h^{k+1,k} \equiv h^{k+1,k}(n)$ defined for $n \in \mathbb{N}$ and $k \in \{0, 1, \dots, n-1\}$ in (6.5) and (6.7) in Section 6.2 above. For such n and k, Proposition 6.11 asserts that $h^{k+1}(n)$ and $h^{k+1,k}(n)$ are identically distributed given $\mathcal{S}_k \equiv \mathcal{S}_k(n) = \sigma(\gamma, m^0, h^j : 1 \leq j \leq k)$, and we now write $h^{k+1,k}(n) = \tilde{h}^{k+1}(n) + \Delta^{k+1}(n)$, where

$$\tilde{h}^1 \equiv \tilde{h}^1(n) := \tau_1 \tilde{Z}^1 \quad \text{and} \quad \Delta^1 \equiv \Delta^1(n) := (\|m^0\|_n - \tau_1)\tilde{Z}^1 + \tilde{\zeta}^1 m^0, \tag{6.20}$$

and

$$\tilde{h}^{k+1} \equiv \tilde{h}^{k+1}(n) := \sum_{\ell=1}^{k} \bar{\alpha}_\ell^k \, h^\ell + \overset{\perp}{\tau}_{k+1} \tilde{Z}^{k+1}, \tag{6.21}$$

$$\begin{aligned}\Delta^{k+1} \equiv \Delta^{k+1}(n) := \sum_{\ell=1}^{k} (\alpha_\ell^k - \bar{\alpha}_\ell^k)\, h^\ell + M_k (M_k^\top M_k)^+ \Biggl(v^{k,k} - \sum_{\ell=1}^{k} \alpha_\ell^k \, v^{k,\ell-1}\Biggr) \\ - \|\overset{\perp}{m}^k\|_n (P_k \tilde{Z}^{k+1}) + (\|\overset{\perp}{m}^k\|_n - \overset{\perp}{\tau}_{k+1}) \tilde{Z}^{k+1} + \tilde{\zeta}^{k+1} \overset{\perp}{m}^k.\end{aligned} \tag{6.22}$$

Recall that $(\tilde{Z}^{k+1}, \tilde{\zeta}^{k+1}) \equiv \bigl(\tilde{Z}^{k+1}(n), \tilde{\zeta}^{k+1}(n)\bigr) \sim N_n(0, I_n) \otimes N(0, 1/n)$ was taken to be independent of $\mathcal{S}_k \equiv \mathcal{S}_k(n)$ in Proposition 6.11, where we also defined $\alpha^k \equiv \alpha^k(n)$ and $v^{k,\ell} \equiv v^{k,\ell}(n)$ for $0 \leq \ell \leq k$.

In the decomposition above, we have defined $\tilde{h}^{k+1}$ in (6.21) to mimic the expression for the limiting Gaussian variable $\bar{G}_{k+1}$ in (6.16). Contrasting the definitions of $h^{k+1,k}$ and $\tilde{h}^{k+1}$ in (6.7) and (6.21) respectively for $k \in \{0, 1, \dots, n-1\}$, we see that the random quantities α^k and $\|\overset{\perp}{m}^k\|_n$ in (6.7) are replaced in (6.21) with the deterministic $\bar{\alpha}^k \in \mathbb{R}^k$ and $\overset{\perp}{\tau}_{k+1} \in (0, \infty)$ from (6.14) and (6.15) respectively; these turn out to be the correct limiting values in Proposition 6.16(i, j) below under the non-degeneracy assumption (A4).

We are now in a position to state the main result of this subsection. To ease notation, we will often suppress the dependence on n of quantities such as $h^k \equiv h^k(n)$, $v^{k,\ell} \equiv v^{k,\ell}(n)$, $\alpha^k \equiv \alpha^k(n)$ and $\Delta^k \equiv \Delta^k(n)$.

Proposition 6.16. *For a sequence of symmetric AMP recursions* (2.1) *satisfying* (A0)–(A5) *as well as* (A4)*, the following hold as $n \to \infty$ for each $k \in \mathbb{N}$:*

(a) $\|\Delta^k\|_{n,r} \xrightarrow{c} 0$;

(b) $\|h^j\|_{n,r} = O_c(1)$ *for* $1 \leq j \leq k$;
$\|m^j\|_{n,r} = O_c(1)$ *for* $0 \leq j \leq k$;

(c) $n^{-1} \sum_{i=1}^n \psi(h_i^1, \dots, h_i^k, \gamma_i) \xrightarrow{c} \mathbb{E}\bigl(\psi(\bar{G}_1, \dots, \bar{G}_k, \bar{\gamma})\bigr)$ *for every* $\psi \in \mathrm{PL}_{k+1}(r)$;

(d) $n^{-1} \sum_{i=1}^n m_i^0 \, \phi(h_i^1, \dots, h_i^k, \gamma_i) \xrightarrow{c} \mathbb{E}\bigl(F_0(\bar{\gamma}) \cdot \phi(\bar{G}_1, \dots, \bar{G}_k, \bar{\gamma})\bigr)$ *for every* $\phi \in \mathrm{PL}_{k+1}(1)$;

(e) $\langle m^{j-1}, m^{\ell-1} \rangle_n \xrightarrow{c} \mathbb{E}(\bar{G}_j \bar{G}_\ell) = \bar{\mathrm{T}}_{j,\ell}$ *for* $1 \leq j, \ell \leq k+1$;

(f) $\langle h^j, m^\ell \rangle_n = \langle h^j, f_\ell(h^\ell, \gamma) \rangle_n \xrightarrow{c} \mathbb{E}\bigl(\bar{G}_j f_\ell(\bar{G}_\ell, \bar{\gamma})\bigr) = \mathbb{E}\bigl(f_\ell'(\bar{G}_\ell, \bar{\gamma})\bigr) \mathbb{E}(\bar{G}_j \bar{G}_\ell) = \bar{b}_\ell \, \bar{\mathrm{T}}_{j,\ell}$ *for* $1 \leq j, \ell \leq k$;
$\langle h^j, m^0 \rangle_n \xrightarrow{c} 0$ *for* $1 \leq j \leq k$;

(g) $b_k = \langle f_k'(h^k, \gamma) \rangle_n \xrightarrow{c} \mathbb{E}\bigl(f_k'(\bar{G}_k, \bar{\gamma})\bigr) = \bar{b}_k$;

(h) $v^{k,\ell}/n = (H_k^\top m^\ell - b_\ell \, M_k^\top m^{\ell-1})/n \xrightarrow{c} 0$ *for* $0 \leq \ell \leq k$;

(i) $\alpha^k \xrightarrow{c} \bar{\alpha}^k$;

(j) $\|\overset{\perp}{m}^k\|_n \xrightarrow{c} \overset{\perp}{\tau}_{k+1} = \mathrm{Var}^{1/2}(\bar{G}_{k+1} \,|\, \bar{G}_1, \dots, \bar{G}_k)$.

Remark 6.17. Under (A4) and the alternative hypotheses of Remark 6.1(a), the assertions (a)–(j) above remain valid if we replace $\xrightarrow{c}$ with $\xrightarrow{p}$ and $O_c(1)$ with $O_p(1)$ throughout.

To establish Proposition 6.16, we proceed by induction on $k \in \mathbb{N}$ and prove the assertions (a)–(i) one at a time (in that order). Here, we will give a technical summary of the inductive argument (which can be read alongside the detailed proof in Section 6.5) to highlight its overall structure and key features. Henceforth, we write $\mathcal{H}_k(\cdots)$ for parts $(\cdots)$ of the inductive hypothesis for $k \in \mathbb{N}$.

$\mathcal{H}_k(e,f)$: These are obtained as direct consequences of the inductive hypotheses $\mathcal{H}_k(c,d)$ by choosing suitable pseudo-Lipschitz functions $\psi \in \mathrm{PL}_{k+1}(2) \subseteq \mathrm{PL}_{k+1}(r)$ that depend on at most three of their $k+1$ arguments. We use $\mathcal{H}_k(d)$ to handle the inner products that feature m^0 and apply $\mathcal{H}_k(c)$ to those that do not. In $\mathcal{H}_k(e)$, the limiting value of $\langle m^{j-1}, m^{\ell-1}\rangle_n = \langle f_{j-1}(h^{j-1},\gamma), f_{\ell-1}(h^{\ell-1},\gamma)\rangle_n$ is shown to be

$$\begin{cases} \mathbb{E}\big(f_{j-1}(\bar{G}_{j-1},\bar{\gamma}) \cdot f_{\ell-1}(\bar{G}_{\ell-1},\bar{\gamma})\big) = \mathbb{E}(\bar{G}_j\bar{G}_\ell) & \text{for } 2 \le j,\ell \le k+1 \\ \mathbb{E}\big(F_0(\bar{\gamma}) \cdot f_{j-1}(\bar{G}_{j-1},\bar{\gamma})\big) = \mathbb{E}(\bar{G}_1\bar{G}_\ell) & \text{for } 1 = j < \ell \le k+1, \end{cases}$$

where the two equalities are drawn from (6.17) and form the basis of the definition of the limiting covariances $\bar{\mathrm{T}}_{j,\ell}$ in (2.6). Moreover, for $1 \le j,\ell \le k$, the identity $\mathbb{E}\big(\bar{G}_j f_\ell(\bar{G}_\ell,\bar{\gamma})\big) = \mathbb{E}\big(f_\ell'(\bar{G}_\ell,\bar{\gamma})\big)\mathbb{E}(\bar{G}_j\bar{G}_\ell)$ in the first line of $\mathcal{H}_k(f)$ comes from (6.19). These identities (6.17) and (6.19) ultimately provide the crucial link between the limiting values of $\langle h^j, m^\ell\rangle_n$ and $b_\ell\,\langle m^{j-1}, m^{\ell-1}\rangle_n$ in $\mathcal{H}_k(h)$.

$\mathcal{H}_k(g,h)$: This is also derived from $\mathcal{H}_k(c)$, but since $f_k'\colon \mathbb{R}^2 \to \mathbb{R}$ need not lie in $\mathrm{PL}_2(r)$, we instead apply the analytic Lemmas 7.10 and 7.14 rather than imitate the proofs of $\mathcal{H}_k(e,f)$. See the proof of Corollary 7.21(b) for a similar argument. $\mathcal{H}_k(h)$ follows immediately from $\mathcal{H}_k(e,f,g)$.

$\mathcal{H}_k(i,j)$: We see from $\mathcal{H}_k(e)$ that the matrices $M_k^\top M_k/n \in \mathbb{R}^{k\times k}$ converge completely to the limiting covariance matrix $\bar{\mathrm{T}}^{[k]} = \mathrm{Cov}(\bar{G}_1,\dots,\bar{G}_k) \in \mathbb{R}^{k\times k}$, which is positive definite under (A4). In $\mathcal{H}_k(i)$, we consider $\alpha^k = (M_k^\top M_k/n)^+(M_k^\top m^k/n) \in \mathbb{R}^k$, a vector of projection coefficients defined in (6.4). It follows from $\mathcal{H}_k(e)$ that $(M_k^\top M_k/n)^+$ and $M_k^\top m^k/n$ converge completely to $(\bar{\mathrm{T}}^{[k]})^{-1}$ and $\bar{\mathrm{T}}^{[k],k+1}$ respectively, and hence that $\alpha^k \xrightarrow{c} (\bar{\mathrm{T}}^{[k]})^{-1}\,\bar{\mathrm{T}}^{[k],k+1} = \bar{\alpha}^k$, as defined in (6.14). For $\mathcal{H}_k(j)$, we recall the definitions at the start of Section 6.2 and write

$$\|\overset{\perp}{m}{}^k\|_n^2 = \|P_k^\perp m^k\|_n^2 = \|m^k\|_n^2 - \|P_k m^k\|_n^2 = \|m^k\|_n^2 - (\alpha^k)^\top(M_k^\top M_k/n)\,\alpha^k.$$

Applying $\mathcal{H}_k(e,i)$ to the individual terms on the right hand side above, we deduce that $\|\overset{\perp}{m}{}^k\|_n^2 \xrightarrow{c} \bar{\mathrm{T}}_{k+1,k+1} - (\bar{\alpha}^k)^\top\bar{\mathrm{T}}^{[k]}\,\bar{\alpha}^k = \overset{\perp}{\tau}{}_{k+1}^2$, as defined in (6.15).

$\mathcal{H}_{k+1}(a)$: It is thanks to the key fact $\mathcal{H}_k(h)$ and the presence of the Onsager term $-b_k m^{k-1}$ in the original AMP recursion (2.1) (and subsequently in (6.7) in Proposition 6.11) that the $\|\cdot\|_{n,r}$ norm of the second term in (6.22) converges completely to 0. Using $\mathcal{H}_k(b,i,j)$ to handle some of the remaining terms in this definition (6.22) of the deviation term Δ^{k+1}, we conclude that $\|\Delta^{k+1}\|_{n,r} \xrightarrow{c} 0$.

$\mathcal{H}_{k+1}(b)$: Using the distributional equality $h^{k+1} \overset{d}{=} h^{k+1,k} = \tilde{h}^{k+1} + \Delta^{k+1} = \sum_{\ell=1}^k \bar{\alpha}_\ell^k\,h^\ell + \overset{\perp}{\tau}_{k+1}\tilde{Z}^{k+1} + \Delta^{k+1}$ from Proposition 6.11 and (6.21, 6.22), we deduce from $\mathcal{H}_{k+1}(a)$ and the inductive hypothesis $\mathcal{H}_k(b)$ that $\|h^{k+1}\|_{n,r} = O_c(1)$. Since $\|\gamma\|_{n,r} = O_c(1)$ by (A1) and f_{k+1} is Lipschitz, this in turn implies that $\|m^{k+1}\|_{n,r} = \|f_k(h^{k+1},\gamma)\|_{n,r} = O_c(1)$.

$\mathcal{H}_{k+1}(c)$: This is the main assertion in Proposition 6.16; by Corollary 7.21(b), it is in fact equivalent to the conclusion (2.8) of Theorem 2.3. We first condition on $\mathcal{S}_k = \sigma(\gamma, m^0, h^j : 1 \le j \le k)$ and appeal to Proposition 6.11, which asserts that for each $n > k$, the conditional distribution of $h^{k+1} \equiv h^{k+1}(n)$ given $\mathcal{S}_k$ is identical to that of $h^{k+1,k} \equiv h^{k+1,k}(n)$ from (6.7). With $h^{k+1,k} = \tilde{h}^{k+1} + \Delta^{k+1}$ in place of h^{k+1} on the left hand side of $\mathcal{H}_{k+1}(c)$, we use $\mathcal{H}_{k+1}(a,b)$ to show that the 'deviation' term $\Delta^{k+1} \equiv \Delta^{k+1}(n)$ from (6.22) has asymptotically negligible effect, so that $h^{k+1,k}$ can in fact be replaced with $\tilde{h}^{k+1}$ in all relevant expressions. In (6.21), $\tilde{h}^{k+1}$ was defined as $\sum_{\ell=1}^k \bar{\alpha}_\ell^k\,h^\ell + \overset{\perp}{\tau}_{k+1}\tilde{Z}^{k+1}$, where $\sum_{\ell=1}^k \bar{\alpha}_\ell^k\,h^\ell$ is a deterministic linear combination of the previous iterates $h^1,\dots,h^k$, and $\overset{\perp}{\tau}_{k+1}\tilde{Z}^{k+1}$ is a new Gaussian variable that has i.i.d. components and is independent of $\mathcal{S}_k$.

In view of this, the proof of $\mathcal{H}_{k+1}(c)$ can be completed in two stages (given by (6.42) and (6.40) below): the influence of the latter Gaussian term can first be understood by appealing to $\mathcal{H}_{k+1}(b)$

and a general concentration result for sums of pseudo-Lipschitz functions of independent Gaussians (Lemma 7.12), before we subsequently reintroduce the randomness in $\gamma, m^0, h^1, \ldots, h^k$ and apply the inductive hypothesis $\mathcal{H}_k(c)$ to account for this. The appearance of the new limiting Gaussian variable $\bar{G}_{k+1}$ on the right hand side of $\mathcal{H}_{k+1}(c)$ (in addition to the existing $\bar{G}_1, \ldots, \bar{G}_k$ from $\mathcal{H}_k(c)$) can be explained through its definition in (6.16), which matches up neatly with the definition (6.21) of $\tilde{h}^{k+1}$ and the two-stage argument we have just outlined; see (6.40) and (6.41) in the proof.

$\mathcal{H}_{k+1}(d)$: The proof of this is similar in spirit to that of $\mathcal{H}_{k+1}(c)$, except that it also makes use of condition (A3). Note also that $\mathcal{H}_{k+1}(d)$ applies only to Lipschitz $\phi\colon \mathbb{R}^{k+2} \to \mathbb{R}$ rather than general $\phi \in \mathrm{PL}_{k+2}(r)$, but this is sufficient for our purposes in the subsequent proofs of $\mathcal{H}_{k+1}(e,f)$.

The proofs we give for $\mathcal{H}_{k+1}(c,d)$ combine aspects of the asymptotic and finite-sample arguments (see Remark 6.3) in the existing AMP literature. Proposition E.1 in Fan (2022) provides the basis for an alternative asymptotic approach, whose details we omit.

6.5 Proofs for Sections 2.1 and 6.1

Proof of Proposition 6.16. Since we are carrying out an asymptotic analysis, we may assume without loss of generality that $n > k$ in the proofs of $\mathcal{H}_k(a, \ldots, i)$ for each $k \in \mathbb{N}$; this enables us to apply the results on conditional distributions from Section 6.2. Note also that we use $T_n, T_{n1}, T_{n1}', T_{n2}$ to refer to different quantities of interest in different parts of the proof. In Lemma 7.2 and Remark 7.3, we state versions of the continuous mapping theorem and Slutsky's lemma for complete convergence, as well the 'arithmetic rules' for o_c and O_c symbols. We will apply these repeatedly in the arguments below, often without further comment or explanation.

First, we prove $\mathcal{H}_1(a,b,c,d)$, which form the base case for the induction.

$\mathcal{H}_1(a)$: By (6.20), $\Delta^1 \equiv \Delta^1(n) = (\|m^0\|_n - \tau_1)\tilde{Z}^1 + \tilde{\zeta}^1 m^0$, where $(\tilde{Z}^1, \tilde{\zeta}^1) \equiv (\tilde{Z}^1(n), \tilde{\zeta}^1(n)) \sim N_n(0, I_n) \otimes N(0, 1/n)$ for each n. Taking $\zeta \sim N(0,1)$, we have $|\tilde{\zeta}^1| \overset{d}{=} n^{-1/2}\,|\zeta| \xrightarrow{c} 0$ by Example 1(a), and $\|\tilde{Z}^1\|_{n,r} = (n^{-1}\sum_{i=1}^n |\tilde{Z}_i^1|^r)^{1/r} \xrightarrow{c} \mathbb{E}(|\zeta|^r)^{1/r} \in (0,\infty)$ by Lemma 7.12 and Proposition 1.2. Moreover, $\big|\|m^0\|_n - \tau_1\big| \xrightarrow{c} 0$ and $\|m^0\|_{n,r} = O_c(1)$ by (A2). Putting everything together, we recall from Remark 7.3 the 'arithmetic rules' (7.1) for o_c and O_c symbols, and conclude using the triangle inequality for $\|\cdot\|_{n,r}$ that

$$\|\Delta^1\|_{n,r} \le \big|\|m^0\|_n - \tau_1\big|\,\|\tilde{Z}^1\|_{n,r} + |\tilde{\zeta}^1|\,\|m^0\|_{n,r} = o_c(1)\,O_c(1) + o_c(1)\,O_c(1) = o_c(1).$$

$\mathcal{H}_1(b)$: Recall from (6.5) in Proposition 6.11 and (6.20) that

$$h^1 \equiv h^1(n) \overset{d}{=}|_{\mathcal{S}_0}\, \tilde{h}^1(n) + \Delta^1(n) = h^{1,0}(n) \equiv h^{1,0} \tag{6.23}$$

for each $n \in \mathbb{N}$, where $\tilde{h}^1 \equiv \tilde{h}^1(n) = \tau_1\tilde{Z}^1$ and $\tilde{Z}^1 \sim N_n(0, I_n)$ is independent of $\mathcal{S}_0 = \sigma(\gamma, m^0)$. Then $\|\Delta^1\|_{n,r} = o_c(1)$ by $\mathcal{H}_1(a)$ and $\|\tilde{h}^1\|_{n,r} = \tau_1\|\tilde{Z}^1\|_{n,r} = O_c(1)$ as in the proof of $\mathcal{H}_1(a)$, so

$$\|h^1\|_{n,r} \overset{d}{=} \|h^{1,0}\|_{n,r} \le \|\tilde{h}^1\|_{n,r} + \|\Delta^1\|_{n,r} = O_c(1) + o_c(1) = O_c(1).$$

We already have $\|m^0\|_{n,r} = O_c(1)$ by (A2). In addition, $\|\gamma\|_{n,r} = (n^{-1}\sum_{i=1}^n |\gamma_i|^r)^{1/r} \xrightarrow{c} \mathbb{E}(|\bar{\gamma}|^r)^{1/r}$ by (A1), so $\|\gamma\|_{n,r} = O_c(1)$. Letting $L' > 0$ be such that the function f_1 in the AMP recursion (2.1) lies in $\mathrm{PL}_2(1, L')$, we have $|f_1(x,y)| \le |f_1(0,0)| + L'(|x| + |y|)$ for all $(x,y) \in \mathbb{R}^2$, so we can apply the triangle inequality for $\|\cdot\|_{n,r}$ to deduce that

$$\|m^1\|_{n,r} = \|f_1(h^1,\gamma)\|_{n,r} \le |f_1(0,0)|\,\|\mathbf{1}_n\|_{n,r} + L'(\|h^1\|_{n,r} + \|\gamma\|_{n,r}) = O_c(1).$$

$\mathcal{H}_1(c)$: For each n, note that $(\gamma, h^1) \overset{d}{=}|_{\mathcal{S}_0} (\gamma, h^{1,0})$ by (6.23) and Lemma 7.6(c). Thus, for each fixed $\psi \in \mathrm{PL}_2(r)$, it follows that $n^{-1}\sum_{i=1}^n \psi(h_i^1, \gamma_i) \overset{d}{=} n^{-1}\sum_{i=1}^n \psi(h_i^{1,0}, \gamma_i) =: T_n$ for each n, so in view of the third bullet point in Remark 6.1, it is enough to show that $T_n \xrightarrow{c} \mathbb{E}\bigl(\psi(\bar{G}_1, \bar{\gamma})\bigr)$ as $n \to \infty$. To this end, we write

$$T_n = \frac{1}{n}\sum_{i=1}^n \psi(\tilde{h}_i^1, \gamma_i) + \frac{1}{n}\sum_{i=1}^n \bigl\{\psi(h_i^{1,0}, \gamma_i) - \psi(\tilde{h}_i^1, \gamma_i)\bigr\} =: T_{n1} + T_{n2}$$

for each n, and aim to prove that $T_{n1} \xrightarrow{c} \mathbb{E}\bigl(\psi(\bar{G}_1, \bar{\gamma})\bigr)$ and $T_{n2} \xrightarrow{c} 0$, which together imply the desired conclusion.

Before proceeding, we briefly describe the techniques that we use to determine the limit of (T_{n1}) and also to prove $\mathcal{H}_1(d)$ and $\mathcal{H}_{k+1}(c,d)$ later on. It is instructive to consider the following two special cases where the claim is easier to establish. If ψ depends only on its first argument, then since $h^1(n) = \tau_1 \tilde{Z}^1(n)$ and $\tilde{Z}^1(n) \sim N_n(0, I_n)$ for each n, the result follows readily from the concentration inequality (7.6) in Lemma 7.12 and the characterisation of complete convergence in Proposition 1.2. On the other hand, if ψ depends only on its second argument, then since $\bigl(\gamma \equiv \gamma(n) : n \in \mathbb{N}\bigr)$ satisfies (A1) by assumption, we can appeal directly to Corollary 7.21(b).

For general $\psi \in \mathrm{PL}_2(r)$, we seek to combine these two different lines of reasoning by exploiting the independence of $\tilde{h}^1(n)$ and $\mathcal{S}_0 \equiv \mathcal{S}_0(n) = \sigma(\gamma, m^0)$ for each n. This allows $\gamma(n)$ and $\tilde{h}^1(n)$ to be handled separately (to a large extent) when we decompose T_{n1} as a sum of $\mathbb{E}(T_{n1} \,|\, \mathcal{S}_0)$ and $T_{n1} - \mathbb{E}(T_{n1} \,|\, \mathcal{S}_0)$ in (6.24) and (6.25) respectively. For the latter, it is helpful to first think of $\gamma(n)$ as being fixed when applying Lemma 7.12 to the Gaussian $\tilde{h}^1$, before subsequently accounting for the randomness of $\gamma(n)$ using (A1).

Define $\Psi \colon \mathbb{R} \to \mathbb{R}$ by $\Psi(y) := \mathbb{E}\bigl(\psi(\tau_1 Z, y)\bigr)$ with $Z \sim N(0,1)$. For each n, since $\tilde{Z}^1 \equiv \tilde{Z}^1(n) \sim N_n(0, I_n)$ is independent of $\mathcal{S}_0 \equiv \mathcal{S}_0(n) = \sigma(\gamma, m^0)$, we deduce from Lemma 7.7 that $\mathbb{E}\bigl(\psi(\tilde{h}_i^1, \gamma_i) \,\big|\, \mathcal{S}_0\bigr) = \mathbb{E}\bigl(\psi(\tau_1 \tilde{Z}_i^1, \gamma_i) \,\big|\, \mathcal{S}_0\bigr) = \Psi(\gamma_i)$ almost surely, for every $1 \le i \le n$. Since $\Psi \in \mathrm{PL}_1(r)$ by Lemma 7.23(b), it follows from (A1) and Corollary 7.21(b) that $n^{-1}\sum_{i=1}^n \Psi(\gamma_i) \xrightarrow{c} \mathbb{E}\bigl(\Psi(\bar{\gamma})\bigr)$ as $n \to \infty$, where $\bar{\gamma} \sim \pi$. A further application of Lemma 7.7 shows that if $\bar{G}_1 \sim N(0, \tau_1^2)$ is independent of $\bar{\gamma}$, then $\mathbb{E}\bigl(\Psi(\bar{\gamma})\bigr) = \mathbb{E}\bigl(\mathbb{E}\{\psi(\bar{G}_1, \bar{\gamma}) \,|\, \bar{\gamma}\}\bigr) = \mathbb{E}\bigl(\psi(\bar{G}_1, \bar{\gamma})\bigr)$, so in summary, we have

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\bigl(\psi(\tilde{h}_i^1, \gamma_i) \,\big|\, \mathcal{S}_0\bigr) = \frac{1}{n}\sum_{i=1}^n \Psi(\gamma_i) \xrightarrow{c} \mathbb{E}\bigl(\Psi(\bar{\gamma})\bigr) = \mathbb{E}\bigl(\psi(\bar{G}_1, \bar{\gamma})\bigr). \tag{6.24}$$

To complete the proof that $T_{n1} \xrightarrow{c} \mathbb{E}\bigl(\psi(\bar{G}_1, \bar{\gamma})\bigr)$, we must therefore show that

$$T'_{n1} := \frac{1}{n}\sum_{i=1}^n \bigl\{\psi(\tilde{h}_i^1, \gamma_i) - \mathbb{E}\bigl(\psi(\tilde{h}_i^1, \gamma_i) \,\big|\, \mathcal{S}_0\bigr)\bigr\} \xrightarrow{c} 0 \tag{6.25}$$

as $n \to \infty$. To this end, let $L > 0$ be such that $\psi \in \mathrm{PL}_2(r, L)$, and for each $y \in \mathbb{R}$, define $\psi_y, \bar{\psi}_y \colon \mathbb{R} \to \mathbb{R}$ by $\psi_y(z) := \psi(\tau_1 z, y)$ and $\bar{\psi}_y(z) := \psi_y(z) - \mathbb{E}\bigl(\psi_y(Z)\bigr)$, where $Z \sim N(0,1)$. Then by Lemma 7.23(a), there exists $K_0 > 0$, depending only on τ_1 and r, such that $\psi_y \in \mathrm{PL}_1(r, K_0 L_y)$ with $L_y := L(1 \vee |y|^{r-1})$. For fixed $n \in \mathbb{N}$ and $y_1, \dots, y_n \in \mathbb{R}$, define $\check{L} \equiv \check{L}(y_1, \dots, y_n) := (L_{y_1}, \dots, L_{y_n})$. Let $r' := r/(r-1) \in (1,2]$ be the Hölder conjugate of r, so that $1/r + 1/r' = 1$, and note that since $\|\cdot\|_{p'} \le \|\cdot\|_p$ for $1 \le p \le p' \le \infty$, we have

$$\begin{aligned}\frac{\|\check{L}\|_\infty}{n^{1/r'}} \le \frac{\|\check{L}\|_2}{n^{1/r'}} \le \frac{\|\check{L}\|_{r'}}{n^{1/r'}} = \|\check{L}\|_{n,r'} = \biggl(\frac{1}{n}\sum_{i=1}^n |L_{y_i}|^{r'}\biggr)^{1/r'} &\le L\biggl(1 + \frac{1}{n}\sum_{i=1}^n |y_i|^r\biggr)^{1/r'} \\ &= L(1 + \|y\|_{n,r}^r)^{1/r'}.\end{aligned} \tag{6.26}$$

By Lemma 7.12, there exists a universal constant $C > 0$ such that if $Z_1, \dots, Z_n \overset{\text{iid}}{\sim} N(0,1)$, then

$$\begin{aligned}
P(n, t, y_1, \dots, y_n) &:= \mathbb{P}\biggl(\biggl|\frac{1}{n}\sum_{i=1}^n \bar{\psi}_{y_i}(Z_i)\biggr| \geq t\biggr) \\
&\leq \exp\biggl(1 - \min\biggl\{\biggl(\frac{nt}{(Cr)^r K_0 \|\breve{L}\|_2}\biggr)^2, \biggl(\frac{nt}{(Cr)^r K_0 \|\breve{L}\|_\infty}\biggr)^{2/r}\biggr\}\biggr) \\
&\leq \exp\biggl(1 - \min\biggl\{\biggl(\frac{n^{1/r}t}{(Cr)^r K_0 \|\breve{L}\|_{n,r'}}\biggr)^2, \biggl(\frac{n^{1/r}t}{(Cr)^r K_0 \|\breve{L}\|_{n,r'}}\biggr)^{2/r}\biggr\}\biggr) \\
&=: E_r(n, t, K_0\|\breve{L}\|_{n,r'}) \equiv E_r\bigl(n, t, K_0\|\breve{L}(y_1, \dots, y_n)\|_{n,r'}\bigr) \qquad (6.27)
\end{aligned}$$

for every $t \geq 0$. Returning to (6.25), we see that

$$\psi(\tilde{h}_i^1, \gamma_i) - \mathbb{E}\bigl(\psi(\tilde{h}_i^1, \gamma_i) \,\big|\, \mathcal{S}_0\bigr) = \psi_{\gamma_i}(\tilde{Z}_i^1) - \mathbb{E}\bigl(\psi_{\gamma_i}(\tilde{Z}_i^1) \,\big|\, \mathcal{S}_0\bigr) = \bar{\psi}_{\gamma_i}(\tilde{Z}_i^1)$$

for all $1 \leq i \leq n$, where the final equality follows from Lemma 7.7 and the fact that $\tilde{Z}^1 \equiv \tilde{Z}^1(n)$ is independent of $\mathcal{S}_0 \equiv \mathcal{S}_0(n) = \sigma(\gamma, m^0)$. We deduce from this and (6.27) that

$$\mathbb{P}(|T'_{n1}| > \varepsilon \,|\, \mathcal{S}_0) = \mathbb{P}\biggl(\biggl|\frac{1}{n}\sum_{i=1}^n \bar{\psi}_{\gamma_i}(\tilde{Z}_i^1)\biggr| \geq \varepsilon \,\biggm|\, \mathcal{S}_0\biggr) = P(n, \varepsilon, \gamma_1, \dots, \gamma_n) \leq E_r\bigl(n, \varepsilon, K_0\breve{L}_0(n)\bigr) \qquad (6.28)$$

for every n and $\varepsilon > 0$, where the second equality is again obtained using Lemma 7.7, and $\breve{L}_0(n) := \|\breve{L}(\gamma_1, \dots, \gamma_n)\|_{n,r'} \leq L(1 + \|\gamma\|_{n,r}^r)^{1/r'} = O_c(1)$ by (6.26) and (A1). Thus, by Proposition 1.2, there exists $\bar{L}_0 \in (0, \infty)$ such that for $n \in \mathbb{N}$, the events $A_0(n) := \{\breve{L}_0(n) \leq \bar{L}_0\} \in \mathcal{S}_0(n)$ satisfy $\sum_{n=1}^\infty \mathbb{P}\bigl(A_0(n)^c\bigr) < \infty$. Moreover, for each n and $\varepsilon > 0$, it follows from (6.28) that

$$\begin{aligned}
\mathbb{P}\bigl(\{|T'_{n1}| > \varepsilon\} \cap A_0(n) \,\big|\, \mathcal{S}_0\bigr) = \mathbb{P}(|T'_{n1}| > \varepsilon \,|\, \mathcal{S}_0)\mathbb{1}_{A_0(n)} &\leq P(n, \varepsilon, \gamma_1, \dots, \gamma_n)\mathbb{1}_{A_0(n)} \\
&\leq E_r\bigl(n, \varepsilon, K_0\breve{L}_0(n)\bigr)\mathbb{1}_{A_0(n)} \leq E_r(n, \varepsilon, K_0\bar{L}_0),
\end{aligned}$$

where we have used the fact that $A_0(n) \in \mathcal{S}_0(n)$ to obtain the first equality above. Recalling the expression for $E_r(n, \varepsilon, K_0\bar{L}_0)$ in (6.27), we see that $\sum_{n=1}^\infty E_r(n, \varepsilon, K_0\bar{L}_0) < \infty$, and hence conclude that for every $\varepsilon > 0$, we have

$$\begin{aligned}
\sum_{n=1}^\infty \mathbb{P}(|T'_{n1}| > \varepsilon) &\leq \sum_{n=1}^\infty \mathbb{P}\bigl(\{|T'_{n1}| > \varepsilon\} \cap A_0(n)\bigr) + \sum_{n=1}^\infty \mathbb{P}\bigl(A_0(n)^c\bigr) \\
&= \sum_{n=1}^\infty \mathbb{E}\bigl\{\mathbb{P}\bigl(\{|T'_{n1}| > \varepsilon\} \cap A_0(n) \,\big|\, \mathcal{S}_0\bigr)\bigr\} + \sum_{n=1}^\infty \mathbb{P}\bigl(A_0(n)^c\bigr) \\
&\leq \sum_{n=1}^\infty E_r(n, \varepsilon, K_0\bar{L}_0) + \sum_{n=1}^\infty \mathbb{P}\bigl(A_0(n)^c\bigr) < \infty, \qquad (6.29)
\end{aligned}$$

which together with Proposition 1.2 implies (6.25). Together with (6.24), this shows that $T_{n1} \xrightarrow{c} \mathbb{E}\bigl(\psi(\bar{G}_1, \bar{\gamma})\bigr)$, as claimed.

Next, we bound $|T_{n2}|$ for each n. Letting $L > 0$ be such that $\psi \in \mathrm{PL}_2(r, L)$, we can apply Lemma 7.24 to see that

$$\begin{aligned}
|T_{n2}| &\leq \frac{1}{n}\sum_{i=1}^n \bigl|\psi(h_i^{1,0}, \gamma_i) - \psi(h_i^{1,0} - \Delta_i^1, \gamma_i)\bigr| \\
&\leq 2^{\frac{r}{2}-1} L\|\Delta^1\|_{n,r}\bigl(1 + \|h^{1,0}\|_{n,r}^{r-1} + \|h^{1,0} - \Delta^1\|_{n,r}^{r-1} + 2\|\gamma\|_{n,r}^{r-1}\bigr) \\
&\lesssim_r L\|\Delta^1\|_{n,r}\bigl(1 + \|h^{1,0}\|_{n,r}^{r-1} + \|\Delta^1\|_{n,r}^{r-1} + \|\gamma\|_{n,r}^{r-1}\bigr), \qquad (6.30)
\end{aligned}$$

where the final bound is obtained using the triangle inequality for $\|\cdot\|_{n,r}$ and the fact that $(a+b)^{r-1} \leq 2^{r-2}(a^{r-1} + b^{r-1})$ for $a, b \geq 0$. Now $\|h^{1,0}\|_{n,r} \overset{d}{=} \|h^1\|_{n,r} = O_c(1)$ by $\mathcal{H}_1(b)$ and $\|\Delta^1\|_{n,r} = o_c(1)$ by

$\mathcal{H}_1(a)$, so $|T_{n2}| = o_c(1)\bigl(1 + O_c(1) + o_c(1)\bigr) = o_c(1)$. We conclude that $n^{-1}\sum_{i=1}^n \psi(h_i^1, \gamma_i) \overset{d}{=} T_n = T_{n1} + T_{n2} \xrightarrow{c} \mathbb{E}\bigl(\psi(\bar{G}_1, \bar{\gamma})\bigr)$, as desired.

$\mathcal{H}_1(d)$: For each n, we have $(\gamma, m^0, h^1) \overset{d}{=} (\gamma, m^0, h^{1,0})$ by (6.23) and Lemma 7.6(c), so for each fixed $\phi \in \mathrm{PL}_2(1)$, it follows that

$$\frac{1}{n}\sum_{i=1}^n m_i^0\, \phi(h_i^1, \gamma_i) \overset{d}{=} \frac{1}{n}\sum_{i=1}^n m_i^0\, \phi(\tilde{h}_i^1, \gamma_i) + \frac{1}{n}\sum_{i=1}^n m_i^0 \bigl\{\phi(h_i^{1,0}, \gamma_i) - \phi(\tilde{h}_i^1, \gamma_i)\bigr\} =: T_{n1} + T_{n2}. \tag{6.31}$$

By similar (and slightly simpler) arguments to those in $\mathcal{H}_1(c)$, we will prove that $T_{n1} \xrightarrow{c} \mathbb{E}\bigl(F_0(\bar{\gamma}) \cdot \phi(\bar{G}_1, \bar{\gamma})\bigr)$ and $T_{n2} \xrightarrow{c} 0$ as $n \to \infty$.

For T_{n1}, define $\Phi\colon \mathbb{R} \to \mathbb{R}$ by $\Phi(y) := \mathbb{E}\bigl(\phi(\tau_1 Z, y)\bigr)$ with $Z \sim N(0,1)$. For each n, recalling once again that $\tilde{Z}^1 \equiv \tilde{Z}^1(n) \sim N_n(0, I_n)$ is independent of $\mathcal{S}_0 \equiv \mathcal{S}_0(n) = \sigma(\gamma, m^0)$, we deduce from Lemma 7.7 that $\mathbb{E}\bigl(\phi(\tilde{h}_i^1, \gamma_i) \bigm| \mathcal{S}_0\bigr) = \Phi(\gamma_i)$ almost surely, for every $1 \le i \le n$. Now since Φ is Lipschitz by Lemma 7.23(b), it follows from (A3) that if $\bar{G}_1 \sim N(0, \tau_1^2)$ is independent of $\bar{\gamma} \sim \pi$, then

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n m_i^0\, \mathbb{E}\bigl(\phi(\tilde{h}_i^1, \gamma_i) \,|\, \mathcal{S}_0\bigr) &= \frac{1}{n}\sum_{i=1}^n m_i^0\, \Phi(\gamma_i) \\ &\xrightarrow{c} \mathbb{E}\bigl(F_0(\bar{\gamma})\Phi(\bar{\gamma})\bigr) = \mathbb{E}\bigl(F_0(\bar{\gamma}) \cdot \mathbb{E}\{\phi(\bar{G}_1, \bar{\gamma}) \,|\, \bar{\gamma}\}\bigr) = \mathbb{E}\bigl(F_0(\bar{\gamma}) \cdot \phi(\bar{G}_1, \bar{\gamma})\bigr).\end{aligned} \tag{6.32}$$

To complete the proof that $T_{n1} \xrightarrow{c} \mathbb{E}\bigl(F_0(\bar{\gamma}) \cdot \phi(\bar{G}_1, \bar{\gamma})\bigr)$, we must therefore show that

$$T_{n1}' := n^{-1}\sum_{i=1}^n m_i^0 \bigl\{\phi(\tilde{h}_i^1, \gamma_i) - \mathbb{E}\bigl(\phi(\tilde{h}_i^1, \gamma_i) \,|\, \mathcal{S}_0\bigr)\bigr\} \xrightarrow{c} 0. \tag{6.33}$$

To this end, let $L > 0$ be such that $\phi \in \mathrm{PL}_2(1, L)$ on $\mathbb{R}$. For $u, y \in \mathbb{R}$, define $\phi_{u,y}\colon \mathbb{R} \to \mathbb{R}$ by $\phi_{u,y}(z) := u\bigl\{\phi(\tau_1 z, y) - \mathbb{E}\bigl(\phi(\tau_1 Z, y)\bigr)\bigr\}$, where $Z \sim N(0,1)$, so that $\phi_{u,y} \in \mathrm{PL}_2(1, L\tau_1|u|)$. Since $\tilde{Z}^1 \sim N_n(0, I_n)$, it follows from (7.12) in Remark 7.13 that for every $v \equiv (v_1, \dots, v_n) \in \mathbb{R}^n$ and $t \ge 0$, we have

$$\begin{aligned}\tilde{P}(n, t, v) := \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \phi_{v_i, y_i}(\tilde{Z}_i^1)\right| \ge t\right) \le \exp\left\{1 - \left(\frac{nt}{CL\tau_1\|v\|_2}\right)^2\right\} \le \exp\left\{1 - \left(\frac{n^{1/2}t}{CL\tau_1\|v\|_n}\right)^2\right\} \\ =: \tilde{E}_r(n, t, L\tau_1\|v\|_n),\end{aligned} \tag{6.34}$$

where $C > 0$ is a suitable universal constant. Recalling once again that $\tilde{Z}^1$ is independent of $\mathcal{S}_0 = \sigma(\gamma, m^0)$, we deduce using Lemma 7.7 that $m_i^0\bigl\{\phi(\tilde{h}_i^1, \gamma_i) - \mathbb{E}\bigl(\phi(\tilde{h}_i^1, \gamma_i) \bigm| \mathcal{S}_0\bigr)\bigr\} = \phi_{m_i^0, \gamma_i}(\tilde{Z}_i^1)$ for all $1 \le i \le n$. Thus, for each n and $\varepsilon > 0$, we have

$$\mathbb{P}(|T_{n1}'| > \varepsilon \,|\, \mathcal{S}_0) = \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \phi_{m_i^0, \gamma_i}(\tilde{Z}_i^1)\right| \ge \varepsilon \,\middle|\, \mathcal{S}_0\right) = \tilde{P}(n, \varepsilon, m^0) \le \tilde{E}_r(n, \varepsilon, L\tau_1\|m^0\|_n). \tag{6.35}$$

Since $\|m^0\|_n \xrightarrow{c} \tau_1$ by (A2), Proposition 1.2 ensures that the events $\tilde{A}_0(n) := \{\|m^0\|_n \le \tau_1 + 1\} \in \mathcal{S}_0(n)$ satisfy $\sum_{n=1}^\infty \mathbb{P}\bigl(\tilde{A}_0(n)^c\bigr) < \infty$. Moreover, for each n and $\varepsilon > 0$, it follows from (6.34) that

$$\begin{aligned}\mathbb{P}\bigl(\{|T_{n1}'| > \varepsilon\} \cap \tilde{A}_0(n) \,|\, \mathcal{S}_0\bigr) = \mathbb{P}(|T_{n1}'| > \varepsilon \,|\, \mathcal{S}_0)\mathbb{1}_{\tilde{A}_0(n)} &\le \tilde{P}(n, \varepsilon, m^0)\mathbb{1}_{\tilde{A}_0(n)} \\ &\le \tilde{E}_r(n, \varepsilon, L\tau_1 C_0)\mathbb{1}_{\tilde{A}_0(n)} \le \tilde{E}_r(n, \varepsilon, L\tau_1 C_0),\end{aligned} \tag{6.36}$$

where we have used the fact that $\tilde{A}_0(n) \in \mathcal{S}_0(n)$ to obtain the first equality above. Recalling the expression for $\tilde{E}_r(n, \varepsilon, C_0)$ in (6.34), we see that $\sum_{n=1}^\infty \tilde{E}_r(n, \varepsilon, L\tau_1 C_0) < \infty$. Thus, for every $\varepsilon > 0$, we conclude as in (6.29) that

$$\sum_{n=1}^{\infty}\mathbb{P}(|T'_{n1}|>\varepsilon)\leq\sum_{n=1}^{\infty}\mathbb{P}\bigl(\{|T'_{n1}|>\varepsilon\}\cap\tilde{A}_0(n)\bigr)+\sum_{n=1}^{\infty}\mathbb{P}\bigl(\tilde{A}_0(n)^c\bigr)$$
$$\leq\sum_{n=1}^{\infty}\tilde{E}_r(n,\varepsilon,L\tau_1C_0)+\sum_{n=1}^{\infty}\mathbb{P}\bigl(\tilde{A}_0(n)^c\bigr)<\infty, \tag{6.37}$$

which implies (6.33) in view of Proposition 1.2, and hence that $T_{n1}\overset{c}{\to}\tilde{\tau}\,\mathbb{E}\bigl(\phi(\bar{G}_1)\bigr)$ in (6.31).

As for T_{n2} in (6.31), let $L>0$ be as above, so that $\phi\in\mathrm{PL}_2(1,L)$. For each n, recalling from (6.23) that $h^{1,0}=\tilde{h}^1+\Delta^1$, we now apply the Cauchy–Schwarz inequality to see that

$$|T_{n2}|\leq\frac{1}{n}\sum_{i=1}^{n}|m_i^0|\,\bigl|\phi(h_i^{1,0},\gamma_i)-\phi(h_i^{1,0}-\Delta_i^1,\gamma_i)\bigr|\leq\frac{L}{n}\sum_{i=1}^{n}|m_i^0|\,|\Delta_i^1|\leq L\|m^0\|_n\|\Delta^1\|_n.$$

Since $\|m^0\|_n\overset{c}{\to}\tau_1$ by (A2) and $\|\Delta^1\|_n\leq\|\Delta^1\|_{n,r}=o_c(1)$ by $\mathcal{H}_1(a)$, we conclude that $T_{n2}=O_c(1)\,o_c(1)=o_c(1)$. This completes the proof of $\mathcal{H}_1(d)$.

Turning to the inductive step, we consider a general $k\in\mathbb{N}$ and suppose that $\mathcal{H}_k(b,c,d)$ have already been established. The assertions $\mathcal{H}_k(e,\dots,j)$ and $\mathcal{H}_{k+1}(a,b,c,d)$ will now be proved, in that order. Note that $\mathrm{PL}_{k+1}(2)\subseteq\mathrm{PL}_{k+1}(r)$ since $r\geq 2$.

$\mathcal{H}_k(e)$: In the case $j=\ell=1$, we have $\bigl|\|m^0\|_n-\tau_1\bigr|\overset{c}{\to}0$ by (A2), so $\langle m^0,m^0\rangle_n\overset{c}{\to}\tau_1^2=\bar{\mathrm{T}}_{1,1}=\mathbb{E}(\bar{G}_1^2)$ by (6.17). Now fix $j,\ell\in\{2,\dots,k+1\}$. Then $\tilde{\psi}_{j\ell}\colon(x_1,\dots,x_k,y)\mapsto f_{j-1}(x_{j-1},y)f_{\ell-1}(x_{\ell-1},y)$ lies in $\mathrm{PL}_{k+1}(2)\subseteq\mathrm{PL}_{k+1}(r)$ by Lemma 7.22 and the fact that $f_{j-1},f_{\ell-1}$ in the AMP recursion (2.1) are Lipschitz by assumption. Thus, by taking $\psi=\tilde{\psi}_{j\ell}$ in $\mathcal{H}_k(c)$, we see that

$$\langle m^{j-1},m^{\ell-1}\rangle_n=\frac{1}{n}\sum_{i=1}^{n}f_{j-1}(h_i^{j-1},\gamma_i)\,f_{\ell-1}(h_i^{\ell-1},\gamma_i)$$
$$\overset{c}{\to}\mathbb{E}\bigl(f_{j-1}(\bar{G}_{j-1},\bar{\gamma})\cdot f_{\ell-1}(\bar{G}_{\ell-1},\bar{\gamma})\bigr)=\bar{\mathrm{T}}_{j,\ell}=\mathbb{E}(\bar{G}_j\bar{G}_\ell),$$

where the final equalities are taken from (6.17). To handle the remaining case where $\{j,\ell\}=\{1,k+1\}$, note that since f_k is Lipschitz, the map $\phi_{k+1}\colon(x_1,\dots,x_k,y)\mapsto f_k(x_k,y)$ lies in $\mathrm{PL}_{k+1}(1)$. Thus, by taking $\phi=\phi_{k+1}$ in $\mathcal{H}_k(d)$, we deduce that

$$\langle m^0,m^k\rangle_n=\frac{1}{n}\sum_{i=1}^{n}m_i^0\,f_k(h_i^k,\gamma_i)\overset{c}{\to}\mathbb{E}\bigl(F_0(\bar{\gamma})\cdot f_k(\bar{G}_k,\bar{\gamma})\bigr)=\bar{\mathrm{T}}_{1,k+1}=\mathbb{E}(\bar{G}_1\bar{G}_{k+1}),$$

where the final equalities are again taken from (6.17).

$\mathcal{H}_k(f)$: This proof is very similar to that of $\mathcal{H}_k(e)$. First fix $1\leq j,\ell\leq k$. By Lemma 7.22, the function $(x_1,\dots,x_k,y)\mapsto x_jf_\ell(x_\ell,y)$ lies in $\mathrm{PL}_{k+1}(2)\subseteq\mathrm{PL}_{k+1}(r)$, so by applying $\mathcal{H}_k(c)$ again, we deduce that

$$\langle h^j,m^\ell\rangle_n=\frac{1}{n}\sum_{i=1}^{n}h_i^j\,f_\ell(h_i^\ell,\gamma_i)\overset{c}{\to}\mathbb{E}\bigl(\bar{G}_jf_\ell(\bar{G}_\ell,\bar{\gamma})\bigr)=\mathbb{E}\bigl(f'_\ell(\bar{G}_\ell,\bar{\gamma})\bigr)\mathbb{E}(\bar{G}_j\bar{G}_\ell)=\bar{b}_\ell\,\bar{\mathrm{T}}_{j,\ell},$$

where the final equalities are taken from (6.19). For the second part of $\mathcal{H}_k(f)$, we fix $1\leq j\leq k$ and apply $\mathcal{H}_k(d)$ with the $\mathrm{PL}_{k+1}(1)$ function $(x_1,\dots,x_k,y)\mapsto x_j$ to see that

$$\langle h^j,m^0\rangle_n=\frac{1}{n}\sum_{i=1}^{n}m_i^0\,h_i^j\overset{c}{\to}\mathbb{E}\bigl(F_0(\bar{\gamma})\bar{G}_j\bigr)=0$$

by the independence of $\bar{G}_j\sim N(0,\tau_j^2)$ and $\bar{\gamma}\sim\pi$.

$\mathcal{H}_k(g)$: In view of Definition 1.1 of complete convergence, it suffices to show that if (β_n) is any sequence of random variables with $\beta_n \overset{d}{=} \langle f_k'(h^k,\gamma)\rangle_n$ for each n, then $\beta_n \xrightarrow{a.s.} \mathbb{E}\big(f_k'(\bar{G}_k,\bar{\gamma})\big)$. For any such sequence (β_n), we first seek to construct a random sequence $\big((\eta_n,\theta_n)\in\mathbb{R}^n\times\mathbb{R}^n : n\in\mathbb{N}\big)$ such that $(\eta_n,\theta_n)\overset{d}{=}\big(h^k(n),\gamma(n)\big)$ for each n, and $\big(\langle f_k'(\eta_n,\theta_n)\rangle_n : n\in\mathbb{N}\big)=(\beta_n : n\in\mathbb{N})$ almost surely as random sequences. This can be done by applying Lemma 7.10, where we take $g_n\colon\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}$ to be the measurable function $(x,y)\mapsto\langle f_k'(x,y)\rangle_n = n^{-1}\sum_{i=1}^n f_k'(x_i,y_i)$ for each n.

Since $(\eta_n,\theta_n)\overset{d}{=}\big(h^k(n),\gamma(n)\big)$ for each n by construction, it follows from the inductive hypothesis $\mathcal{H}_k(c)$ that $n^{-1}\sum_{i=1}^n\varphi(\eta_{ni},\theta_{ni})\overset{d}{=}n^{-1}\sum_{i=1}^n\varphi(h_i^k,\gamma_i)\xrightarrow{c}\mathbb{E}\big(\varphi(\bar{G}_k,\bar{\gamma})\big)$ for every $\varphi\in\mathrm{PL}_2(r)$, where $(\bar{G}_k,\bar{\gamma})\sim N(0,\tau_k^2)\otimes\pi =: \bar{\mu}^k$. Consequently, denoting by $\tilde{\mu}_n^k := \nu_n(\eta_n,\theta_n) = n^{-1}\sum_{i=1}^n\delta_{(\eta_{ni},\theta_{ni})}$ the joint empirical distribution of the components of η_n and θ_n for each n, we deduce using Corollary 7.21(a) that $d_r(\tilde{\mu}_n^k,\bar{\mu}^k)\xrightarrow{a.s.}0$, and hence that $(\tilde{\mu}_n^k)$ converges weakly to $\bar{\mu}^k$ with probability 1. By (A5), f_k' is bounded, Borel measurable and continuous $\bar{\mu}^k$-almost everywhere, so we may now apply Lemma 7.14 to conclude that $\beta_n=\langle f_k'(\eta_n,\theta_n)\rangle_n=\int_{\mathbb{R}^2}f_k'\,d\tilde{\mu}_n^k\to\int_{\mathbb{R}^2}f_k'\,d\bar{\mu}^k=\mathbb{E}\big(f_k'(\bar{G}_k,\bar{\gamma})\big)$ almost surely. This completes the proof of $\mathcal{H}_k(g)$.

$\mathcal{H}_k(h)$: For $1\le\ell\le k$, it follows from $\mathcal{H}_k(e,f,g)$ that

$$v_j^{k,\ell}/n=\langle h^j,m^\ell\rangle_n-b_\ell\,\langle m^{j-1},m^{\ell-1}\rangle_n\xrightarrow{c}\bar{b}_\ell\,\bar{\mathrm{T}}_{j,\ell}-\bar{b}_\ell\,\bar{\mathrm{T}}_{j,\ell}=0$$

for all $1\le j\le k$. For $\ell=0$, we have $m^{-1}=0$ by definition, so $v_j^{k,0}/n=\langle h^j,m^0\rangle_n\xrightarrow{c}0$ for all $1\le j\le k$ by the second part of $\mathcal{H}_k(f)$.

$\mathcal{H}_k(i)$: Recall from (6.4) that $\alpha^k=(M_k^\top M_k/n)^+(M_k^\top m^k/n)\in\mathbb{R}^k$. It follows from $\mathcal{H}_k(e)$ that $(M_k^\top M_k/n)_{j\ell}=\langle m^{j-1},m^{\ell-1}\rangle_n\xrightarrow{c}\bar{\mathrm{T}}_{j,\ell}$ and $(M_k^\top m^k/n)_j=\langle m^{j-1},m^k\rangle_n\xrightarrow{c}\bar{\mathrm{T}}_{j,k+1}$ for all $1\le j,\ell\le k$. In the notation of Section 6.4, this means that $M_k^\top m^k/n\xrightarrow{c}\bar{\mathrm{T}}^{[k],k+1}\in\mathbb{R}^k$ and $M_k^\top M_k/n\xrightarrow{c}\bar{\mathrm{T}}^{[k]}\in\mathbb{R}^{k\times k}$. Under (A4), Lemma 2.2 ensures that $\bar{\mathrm{T}}^{[k]}$ is positive definite and hence invertible, we now apply the continuous mapping theorem for complete convergence (Lemma 7.2) to deduce that

$$\alpha^k=(M_k^\top M_k/n)^+(M_k^\top m^k/n)\xrightarrow{c}\big(\bar{\mathrm{T}}^{[k]}\big)^{-1}\,\bar{\mathrm{T}}^{[k],k+1}=\bar{\alpha}^k,$$

as defined in (6.14).

$\mathcal{H}_k(j)$: Recalling (6.4) as well as the definitions at the start of Section 6.2, we can write

$$\|\overset{\perp}{m}^k\|_n^2=\|P_k^\perp m^k\|_n^2=\|m^k\|_n^2-\|P_k m^k\|_n^2=\|m^k\|_n^2-(\alpha^k)^\top(M_k^\top M_k/n)\,\alpha^k.$$

Now $\|m^k\|_n^2\xrightarrow{c}\bar{\mathrm{T}}_{k+1,k+1}=\tau_{k+1}^2$ and $M_k^\top M_k/n\xrightarrow{c}\bar{\mathrm{T}}^{[k]}\in\mathbb{R}^{k\times k}$ by $\mathcal{H}_k(e)$, and $\alpha^k\xrightarrow{c}\bar{\alpha}^k\in\mathbb{R}^k$ by $\mathcal{H}_k(i)$, so

$$\|\overset{\perp}{m}^k\|_n^2=\|m^k\|_n^2-(\alpha^k)^\top(M_k^\top M_k/n)\,\alpha^k\xrightarrow{c}\bar{\mathrm{T}}_{k+1,k+1}-(\bar{\alpha}^k)^\top\bar{\mathrm{T}}^{[k]}\,\bar{\alpha}^k=\overset{\perp}{\tau}{}_{k+1}^2,$$

as defined in (6.15).

$\mathcal{H}_{k+1}(a)$: Denote by $R_{n1},\dots,R_{n5}$ the individual summands (in the order in which they appear) in the definition (6.22) of $\Delta^{k+1}\equiv\Delta^{k+1}(n)\in\mathbb{R}^n$. To establish that $\|\Delta^{k+1}\|_{n,r}\xrightarrow{c}0$, it suffices to show that $\|R_{ns}\|_{n,r}\xrightarrow{c}0$ for $s=1,\dots,5$. Observe first that since $\alpha^k\xrightarrow{c}\bar{\alpha}^k\in\mathbb{R}^k$ by $\mathcal{H}_k(i)$ and $\|h^\ell\|_{n,r}=O_c(1)$ for all $1\le\ell\le k$ by $\mathcal{H}_k(b)$, we have $\|R_{n1}\|_{n,r}\le\sum_{\ell=1}^k|\alpha_\ell^k-\bar{\alpha}_\ell^k|\,\|h^\ell\|_{n,r}=\sum_{\ell=1}^k o_c(1)\,O_c(1)=o_c(1)$.

As for R_{n2}, we know from $\mathcal{H}_k(e)$ that $(M_k^\top M_k/n)_{j\ell}=\langle m^{j-1},m^{\ell-1}\rangle_n\xrightarrow{c}\bar{\mathrm{T}}_{j,\ell}$ for all $1\le j,\ell\le k$, so $M_k^\top M_k/n\xrightarrow{c}\bar{\mathrm{T}}^{[k]}\in\mathbb{R}^{k\times k}$, which is positive definite by Lemma 2.2. We can now apply the continuous mapping theorem for complete convergence (Lemma 7.2) to deduce that $(M_k^\top M_k/n)^+\xrightarrow{c}(\bar{\mathrm{T}}^{[k]})^{-1}$; see $\mathcal{H}_k(i)$ above for a similar argument. By $\mathcal{H}_k(h)$, we have $v^{k,\ell}/n\xrightarrow{c}0\in\mathbb{R}^k$ for all $0\le\ell\le k$, so

$$\tilde{w}^k\equiv\tilde{w}^k(n):=(M_k^\top M_k)^+\Bigg(v^{k,k}-\sum_{\ell=1}^k\alpha_\ell^k\,v^{k,\ell-1}\Bigg)\xrightarrow{c}0,$$

formally by Slutsky's lemma for complete convergence (Lemma 7.2). Since $\|m^{\ell-1}\|_{n,r} = O_c(1)$ for $1 \leq \ell \leq k$ by $\mathcal{H}_k(b)$, we have $\|R_{n2}\|_{n,r} = \|M_k \tilde{w}^k\|_{n,r} \leq \sum_{\ell=1}^k |\tilde{w}_\ell^k|\,\|m^{\ell-1}\|_{n,r} = \sum_{\ell=1}^k o_c(1)\,O_c(1) = o_c(1)$.

Turning to R_{n3} and introducing $\xi_1, \ldots, \xi_k \overset{\text{iid}}{\sim} N(0,1)$, we see from Lemma 6.18 that $\|P_k \tilde{Z}^{k+1}\|_{n,r}$ is stochastically dominated by $\sum_{i=1}^k |\xi_i|/n^{1/r}$ for each $n > k$. By Example 1(a), $\sum_{i=1}^k |\xi_i|/n^{1/r} \leq k \max_{1\leq i\leq k} |\xi_i|/n^{1/r} = o_c(1)$, so $\|P_k \tilde{Z}^{k+1}\|_{n,r} \xrightarrow{c} 0$. Since $\|\overset{\perp}{m}{}^k\|_n \xrightarrow{c} \overset{\perp}{\tau}_{k+1} \in (0,\infty)$ by $\mathcal{H}_k(j)$, we deduce that $\|R_{n3}\|_{n,r} = \|\overset{\perp}{m}{}^k\|_n \|P_k \tilde{Z}^{k+1}\|_{n,r} = o_c(1)$.

For the remaining summands R_{n4} and R_{n5}, the arguments are similar to those in the proof of $\mathcal{H}_1(a)$. Recall that $(\tilde{Z}^{k+1}, \tilde{\zeta}^{k+1}) \equiv \bigl(\tilde{Z}^{k+1}(n), \tilde{\zeta}^{k+1}(n)\bigr) \sim N_n(0, I_n) \otimes N(0, 1/n)$. Introducing $\zeta \sim N(0,1)$, we have $|\tilde{\zeta}^{k+1}| \overset{d}{=} n^{-1/2}|\zeta| \xrightarrow{c} 0$ by Example 1(a), and $\|\tilde{Z}^{k+1}\|_{n,r} = (n^{-1}\sum_{i=1}^n |\tilde{Z}_i^{k+1}|^r)^{1/r} \xrightarrow{c} \mathbb{E}(|\zeta|^r)^{1/r} \in (0,\infty)$ by Lemma 7.12 and Proposition 1.2. By $\mathcal{H}_k(j)$, we have $\|\overset{\perp}{m}{}^k\|_n - \overset{\perp}{\tau}_{k+1} = o_c(1)$. Moreover, $\alpha^k = \bar{\alpha}^k + o_c(1) = O_c(1)$ by $\mathcal{H}_k(i)$ and $\|m^\ell\|_{n,r} = O_c(1)$ for $0 \leq \ell \leq k$ by $\mathcal{H}_k(b)$, so it follows from (6.4) that

$$\|\overset{\perp}{m}{}^k\|_{n,r} = \|(I - P_k)\,m^k\|_{n,r} \leq \|m^k\|_{n,r} + \sum_{\ell=1}^k |\alpha_\ell^k|\,\|m^{\ell-1}\|_{n,r} = O_c(1) + \sum_{\ell=1}^k O_c(1)\,O_c(1) = O_c(1).$$

Putting everything together, we see that

$$\|R_{n4}\|_{n,r} + \|R_{n5}\|_{n,r} = \bigl|\|\overset{\perp}{m}{}^k\|_n - \overset{\perp}{\tau}_{k+1}\bigr|\,\|\tilde{Z}^{k+1}\|_{n,r} + |\tilde{\zeta}^{k+1}|\,\|\overset{\perp}{m}{}^k\|_{n,r} = o_c(1)\,O_c(1) + o_c(1)\,O_c(1) = o_c(1).$$

We have now shown that $\|R_{ns}\|_{n,r} \xrightarrow{c} 0$ for $s = 1, \ldots, 5$, so $\|\Delta^{k+1}\|_{n,r} \leq \sum_{s=1}^5 \|R_{ns}\|_{n,r} \xrightarrow{c} 0$.

$\mathcal{H}_{k+1}(b)$: By the inductive hypothesis $\mathcal{H}_k(b)$, we have $\|h^j\|_{n,r} = O_c(1)$ for all $1 \leq j \leq k$ and $\|m^j\|_{n,r} = O_c(1)$ for all $0 \leq j \leq k$. Now let $j = k+1$. For each integer $n > k$, recall from (6.7) in Proposition 6.11 and (6.21) that

$$h^{k+1}(n) \overset{d}{=}|_{\mathcal{S}_k}\, h^{k+1,k}(n) = \tilde{h}^{k+1}(n) + \Delta^{k+1}(n) = \sum_{\ell=1}^k \bar{\alpha}_\ell^k\, h^\ell(n) + \overset{\perp}{\tau}_{k+1}\tilde{Z}^{k+1}(n) + \Delta^{k+1}(n), \tag{6.38}$$

where the deterministic $\bar{\alpha}^k \in \mathbb{R}^k$ is taken from (6.14) and $\tilde{Z}^{k+1} \equiv \tilde{Z}^{k+1}(n) \sim N_n(0, I_n)$ is independent of $\mathcal{S}_k \equiv \mathcal{S}_k(n) = \sigma(\gamma, m^0, h^1, \ldots, h^k)$. Then $\|\Delta^{k+1}\|_{n,r} = o_c(1)$ by $\mathcal{H}_{k+1}(a)$ and $\|\tilde{Z}^{k+1}\|_{n,r} = O_c(1)$, as in the last part of the proof of $\mathcal{H}_{k+1}(a)$ above. It follows from this and $\mathcal{H}_k(b)$ that

$$\|h^{k+1}\|_{n,r} \overset{d}{=} \|h^{k+1,k}\|_{n,r} \leq \sum_{\ell=1}^k |\bar{\alpha}_\ell^k|\,\|h^\ell\|_{n,r} + \|\Delta^{k+1}\|_{n,r} = O_c(1) + o_c(1) = O_c(1).$$

In addition, letting $L' > 0$ be such that f_{k+1} is L'-Lipschitz, we can argue as in the proof of $\mathcal{H}_1(b)$ to deduce that

$$\|m^{k+1}\|_{n,r} = \|f_{k+1}(h^{k+1}, \gamma)\|_{n,r} \leq |f_{k+1}(0,0)| + L'(\|h^{k+1}\|_{n,r} + \|\gamma\|_{n,r}) = O_c(1).$$

$\mathcal{H}_{k+1}(c)$: We again make use of the distributional equality (6.38), which together with Lemma 7.6(c) implies that $(\gamma, h^1, \ldots, h^k, h^{k+1}) \overset{d}{=}|_{\mathcal{S}_k} (\gamma, h^1, \ldots, h^k, h^{k+1,k})$ for each integer $n > k$. Thus, for any fixed $\psi \in \mathrm{PL}_{k+2}(r)$, it follows that $n^{-1}\sum_{i=1}^n \psi(h_i^1, \ldots, h_i^k, h_i^{k+1}, \gamma_i) \overset{d}{=} n^{-1}\sum_{i=1}^n \psi(h_i^1, \ldots, h_i^k, h_i^{k+1,k}, \gamma_i) =: T_n$ for each such n, so it suffices to show that $T_n \xrightarrow{c} \mathbb{E}\bigl(\psi(\bar{G}_1, \ldots, \bar{G}_k, \bar{G}_{k+1}, \bar{\gamma})\bigr)$. We decompose T_n as

$$\frac{1}{n}\sum_{i=1}^n \psi(h_i^1, \ldots, h_i^k, \tilde{h}_i^{k+1}, \gamma_i) + \frac{1}{n}\sum_{i=1}^n \{\psi(h_i^1, \ldots, h_i^k, h_i^{k+1,k}, \gamma_i) - \psi(h_i^1, \ldots, h_i^k, \tilde{h}_i^{k+1}, \gamma_i)\} =: T_{n1} + T_{n2} \tag{6.39}$$

for each $n > k$, and seek to establish that $T_{n1} \xrightarrow{c} \mathbb{E}\bigl(\psi(\bar{G}_1, \ldots, \bar{G}_k, \bar{G}_{k+1}, \bar{\gamma})\bigr)$ and $T_{n2} \xrightarrow{c} 0$ by imitating and extending the analogous arguments in the proof of $\mathcal{H}_1(c)$.

For T_{n1}, define $\Psi\colon \mathbb{R}^{k+1} \to \mathbb{R}$ by $\Psi(x_1, \dots, x_k, y) := \mathbb{E}\bigl\{\psi\bigl(x_1, \dots, x_k, \sum_{\ell=1}^k \bar{\alpha}_\ell^k\, x_\ell + \overset{\perp}{\tau}_{k+1} Z, y\bigr)\bigr\}$ with $Z \sim N(0,1)$. For each integer $n > k$, since $\tilde{Z}^{k+1} \equiv \tilde{Z}^{k+1}(n) \sim N_n(0, I_n)$ is independent of $\mathcal{S}_k \equiv \mathcal{S}_k(n) = \sigma(\gamma, m^0, h^1, \dots, h^k)$, we deduce from (6.38) and Lemma 7.7 that

$$\mathbb{E}\bigl(\psi(h_i^1, \dots, h_i^k, \tilde{h}_i^{k+1}, \gamma_i) \,\big|\, \mathcal{S}_k\bigr) = \mathbb{E}\bigl\{\psi\bigl(h_i^1, \dots, h_i^k, \textstyle\sum_{\ell=1}^k \bar{\alpha}_\ell^k\, h_i^\ell + \overset{\perp}{\tau}_{k+1} \tilde{Z}_i^{k+1}, \gamma_i\bigr) \,\big|\, \mathcal{S}_k\bigr\} = \Psi(h_i^1, \dots, h_i^k, \gamma_i)$$

almost surely, for every $1 \le i \le n$. Now since $\Psi \in \mathrm{PL}_{k+1}(r)$ by Lemma 7.23(b), it follows from the inductive hypothesis $\mathcal{H}_k(c)$ that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\bigl(\psi(h_i^1, \dots, h_i^k, \tilde{h}_i^{k+1}, \gamma_i) \,\big|\, \mathcal{S}_k\bigr) = \frac{1}{n}\sum_{i=1}^n \Psi(h_i^1, \dots, h_i^k, \gamma_i) \xrightarrow{c} \mathbb{E}\bigl(\Psi(\bar{G}_1, \dots, \bar{G}_k, \bar{\gamma})\bigr), \tag{6.40}$$

where the equality above holds almost surely for each $n > k$. Taking $\overset{\perp}{\zeta}_{k+1} \sim N(0,1)$ to be independent of $\bar{G}_{[k]} := (\bar{G}_1, \dots, \bar{G}_k)$ and $\bar{\gamma}$, we apply Lemma 7.7 again to see that

$$\begin{aligned}\mathbb{E}\bigl(\Psi(\bar{G}_1, \dots, \bar{G}_k, \bar{\gamma})\bigr) &= \mathbb{E}\bigl(\mathbb{E}\bigl\{\psi\bigl(\bar{G}_1, \dots, \bar{G}_k, \textstyle\sum_{\ell=1}^k \bar{\alpha}_\ell^k\, \bar{G}_\ell + \overset{\perp}{\tau}_{k+1}\overset{\perp}{\zeta}_{k+1}, \bar{\gamma}\bigr) \,\big|\, \bar{G}_{[k]}, \bar{\gamma}\bigr\}\bigr) \qquad (6.41)\\ &= \mathbb{E}\bigl\{\psi\bigl(\bar{G}_1, \dots, \bar{G}_k, \textstyle\sum_{\ell=1}^k \bar{\alpha}_\ell^k\, \bar{G}_\ell + \overset{\perp}{\tau}_{k+1}\overset{\perp}{\zeta}_{k+1}, \bar{\gamma}\bigr)\bigr\} = \mathbb{E}\bigl(\psi(\bar{G}_1, \dots, \bar{G}_k, \bar{G}_{k+1}, \bar{\gamma})\bigr),\end{aligned}$$

where the final equality follows from the definition of $\bar{G}_{k+1}$ in (6.16). To complete the proof that $T_{n1} \xrightarrow{c} \mathbb{E}\bigl(\psi(\bar{G}_1, \dots, \bar{G}_k, \bar{G}_{k+1}, \bar{\gamma})\bigr)$, we must therefore show that

$$T_{n1}' := \frac{1}{n}\sum_{i=1}^n \bigl\{\psi(h_i^1, \dots, h_i^k, \tilde{h}_i^{k+1}, \gamma_i) - \mathbb{E}\bigl(\psi(h_i^1, \dots, h_i^k, \tilde{h}_i^{k+1}, \gamma_i) \,\big|\, \mathcal{S}_k\bigr)\bigr\} \xrightarrow{c} 0. \tag{6.42}$$

To this end, let $L > 0$ be such that $\psi \in \mathrm{PL}_{k+2}(r, L)$, and for each $v \equiv (x_1, \dots, x_k, y) \in \mathbb{R}^{k+1}$, define $\psi_v, \bar{\psi}_v \colon \mathbb{R} \to \mathbb{R}$ by $\psi_v(z) := \psi\bigl(x_1, \dots, x_k, \sum_{\ell=1}^k \bar{\alpha}_\ell^k\, x_\ell + \overset{\perp}{\tau}_{k+1} z, y\bigr)$ and $\bar{\psi}_v(z) := \psi_v(z) - \mathbb{E}\bigl(\psi_v(Z)\bigr)$, where $Z \sim N(0,1)$. Then by Lemma 7.23(a), there exists $K_k > 0$, depending only on the deterministic $\bar{\alpha}^k \equiv (\bar{\alpha}_1^k, \dots, \bar{\alpha}_k^k)$, $\overset{\perp}{\tau}_{k+1}$ and r, such that $\psi_v \in \mathrm{PL}_1(r, K_k L_{\|v\|})$ with $L_{\|v\|} = L(1 \vee \|v\|^{r-1})$. For a fixed integer $n > k$ and $v^{(1)}, \dots, v^{(n)} \in \mathbb{R}^{k+1}$, define $\check{L} \equiv \check{L}\bigl(v^{(1)}, \dots, v^{(n)}\bigr) := \bigl(L_{\|v^{(1)}\|}, \dots, L_{\|v^{(n)}\|}\bigr)$. Let $r' = r/(r-1) \in (1,2]$ be as in the proof of $\mathcal{H}_1(c)$, so that $1/r + 1/r' = 1$, and note that since $\|\cdot\|_{p'} \le \|\cdot\|_p$ for $1 \le p \le p' \le \infty$, we have

$$\frac{\|\check{L}\|_\infty}{n^{1/r'}} \le \frac{\|\check{L}\|_2}{n^{1/r'}} \le \frac{\|\check{L}\|_{r'}}{n^{1/r'}} = \|\check{L}\|_{n,r'} = \biggl(\frac{1}{n}\sum_{i=1}^n \bigl(L_{\|v^{(i)}\|}\bigr)^{r'}\biggr)^{1/r'} \le L\biggl(1 + \frac{1}{n}\sum_{i=1}^n \|v^{(i)}\|^r\biggr)^{1/r'}. \tag{6.43}$$

By Lemma 7.12, it follows as in (6.27) that there exists a universal constant $C > 0$ such that if $Z_1, \dots, Z_n \overset{\mathrm{iid}}{\sim} N(0,1)$, then

$$\begin{aligned}P\bigl(n, t, v^{(1)}, \dots, v^{(n)}\bigr) &:= \mathbb{P}\biggl(\biggl|\frac{1}{n}\sum_{i=1}^n \bar{\psi}_{v^{(i)}}(Z_i)\biggr| \ge t\biggr)\\ &\le \exp\biggl(1 - \min\biggl\{\biggl(\frac{nt}{(Cr)^r K_k \|\check{L}\|_2}\biggr)^2, \biggl(\frac{nt}{(Cr)^r K_k \|\check{L}\|_\infty}\biggr)^{2/r}\biggr\}\biggr)\\ &\le \exp\biggl(1 - \min\biggl\{\biggl(\frac{n^{1/r}t}{(Cr)^r K_k \|\check{L}\|_{n,r'}}\biggr)^2, \biggl(\frac{n^{1/r}t}{(Cr)^r K_k \|\check{L}\|_{n,r'}}\biggr)^{2/r}\biggr\}\biggr)\\ &= E_r(n, t, K_k\|\check{L}\|_{n,r'}) \equiv E_r\Bigl(n, t, K_k\bigl\|\check{L}(v^{(1)}, \dots, v^{(n)})\bigr\|_{n,r'}\Bigr) \qquad (6.44)\end{aligned}$$

for every $t \ge 0$. Next, define the $\mathcal{S}_k$-measurable vectors $v_k^{(i)} := (h_i^1, \dots, h_i^k, \gamma_i)$ for $1 \le i \le n$. Returning to (6.42) and recalling (6.38), we see that

$$\begin{aligned}\psi(h_i^1, \dots, h_i^k, \tilde{h}_i^{k+1}, \gamma_i) - \mathbb{E}\bigl(\psi(h_i^1, \dots, h_i^k, \tilde{h}_i^{k+1}, \gamma_i) \,\big|\, \mathcal{S}_k\bigr) &= \psi_{v_k^{(i)}}(\tilde{Z}_i^{k+1}) - \mathbb{E}\bigl(\psi_{v_k^{(i)}}(\tilde{Z}_i^{k+1}) \,\big|\, \mathcal{S}_k\bigr)\\ &= \bar{\psi}_{v_k^{(i)}}(\tilde{Z}_i^{k+1})\end{aligned}$$

for all $1 \le i \le n$, where the final equality follows from Lemma 7.7 and the fact that $\tilde{Z}^{k+1} \equiv \tilde{Z}^{k+1}(n)$ is independent of $\mathcal{S}_k \equiv \mathcal{S}_k(n)$. We deduce from this and (6.44) that

$$\begin{aligned}\mathbb{P}(|T'_{n1}| > \varepsilon \,|\, \mathcal{S}_k) = \mathbb{P}\biggl(\biggl|\frac{1}{n}\sum_{i=1}^{n} \bar{\psi}_{\upsilon_k^{(i)}}(\tilde{Z}_i^{k+1})\biggr| \ge \varepsilon \biggm| \mathcal{S}_k\biggr) &= P\bigl(n, \varepsilon, \upsilon_k^{(1)}, \dots, \upsilon_k^{(n)}\bigr) \\ &\le E_r\Bigl(n, \varepsilon, K_k \bigl\|\breve{L}\bigl(\upsilon_k^{(1)}, \dots, \upsilon_k^{(n)}\bigr)\bigr\|_{n,r'}\Bigr) \qquad (6.45)\end{aligned}$$

for every $n > k$ and $\varepsilon > 0$, where the second equality is again obtained using Lemma 7.7. Now by (6.43) and Hölder's inequality, which ensures that $\|\cdot\| \equiv \|\cdot\|_2 \le (k+1)^{\frac{1}{2}-\frac{1}{r}}\,\|\cdot\|_r$ on $\mathbb{R}^{k+1}$, we have

$$\begin{aligned}\breve{L}_k(n) := \bigl\|\breve{L}\bigl(\upsilon_k^{(1)}, \dots, \upsilon_k^{(n)}\bigr)\bigr\|_{n,r'} &\le L\biggl(1 + \frac{1}{n}\sum_{i=1}^{n} \|\upsilon_k^{(i)}\|^r\biggr)^{1/r'} \\ &\lesssim_{k,r} L\biggl\{1 + \frac{1}{n}\sum_{i=1}^{n}\sum_{\ell=1}^{k}\bigl(|h_i^\ell|^r + |\gamma_i|^r\bigr)\biggr\}^{1/r'} \\ &= L\biggl(1 + \sum_{\ell=1}^{k}\bigl(\|h^\ell\|_{n,r}^r + \|\gamma\|_{n,r}^r\bigr)\biggr)^{1/r'} \qquad (6.46)\end{aligned}$$

for each $n > k$. Since $\|h^\ell\|_{n,r} = O_c(1)$ for $1 \le \ell \le k$ by $\mathcal{H}_k(b)$ and $\|\gamma\|_{n,r} = O_c(1)$ by (A1), this means that $\breve{L}_k(n) = O_c(1)$. Thus, by Proposition 1.2, there exists $\bar{L}_k \in (0, \infty)$ such that for integers $n > k$, the events $A_k(n) := \{\breve{L}_k(n) \le \bar{L}_k\} \in \mathcal{S}_k(n)$ satisfy $\sum_{n=k+1}^{\infty} \mathbb{P}\bigl(A_k(n)^c\bigr) < \infty$. Moreover, for each $n > k$ and $\varepsilon > 0$, it follows from (6.45) that

$$\begin{aligned}\mathbb{P}\bigl(\{|T'_{n1}| > \varepsilon\} \cap A_k(n) \,\big|\, \mathcal{S}_k\bigr) = \mathbb{P}(|T'_{n1}| > \varepsilon \,|\, \mathcal{S}_k)\mathbb{1}_{A_k(n)} &\le P\bigl(n, \varepsilon, \upsilon_k^{(1)}, \dots, \upsilon_k^{(n)}\bigr)\mathbb{1}_{A_k(n)} \\ &\le E_r\bigl(n, \varepsilon, K_k\breve{L}_k(n)\bigr)\mathbb{1}_{A_k(n)} \le E_r(n, \varepsilon, K_k\bar{L}_k),\end{aligned}$$

where we have used the fact that $A_k(n) \in \mathcal{S}_k(n)$ to obtain the first equality above. Recalling the expression for $E_r(n, \varepsilon, K_k\bar{L}_k)$ in (6.44), we see that $\sum_{n=k+1}^{\infty} E_r(n, \varepsilon, K_k\bar{L}_k) < \infty$, and hence conclude as in (6.29) that for every $\varepsilon > 0$, we have

$$\begin{aligned}\sum_{n=k+1}^{\infty} \mathbb{P}(|T'_{n1}| > \varepsilon) &\le \sum_{n=k+1}^{\infty} \mathbb{P}\bigl(\{|T'_{n1}| > \varepsilon\} \cap A_k(n)\bigr) + \sum_{n=k+1}^{\infty} \mathbb{P}\bigl(A_k(n)^c\bigr) \\ &\le \sum_{n=k+1}^{\infty} E_r(n, \varepsilon, K_k\bar{L}_k) + \sum_{n=k+1}^{\infty} \mathbb{P}\bigl(A_k(n)^c\bigr) < \infty,\end{aligned}$$

which implies (6.42) in view of Proposition 1.2. Together with (6.40) and (6.41), this shows that $T_{n1} \xrightarrow{c} \mathbb{E}\bigl(\psi(\bar{G}_1, \dots, \bar{G}_k, \bar{G}_{k+1}, \bar{\gamma})\bigr)$ in (6.39), as claimed.

The final step in the proof of $\mathcal{H}_{k+1}(c)$ is to show that $T_{n2} \xrightarrow{c} 0$ in (6.39). Letting $L > 0$ be such that $\psi \in \mathrm{PL}_{k+2}(r, L)$, we can apply Lemma 7.24 as in (6.30) to see that

$$\begin{aligned}|T_{n2}| &\le \frac{1}{n}\sum_{i=1}^{n}\bigl|\psi(h_i^1, \dots, h_i^k, h_i^{k+1,k}, \gamma_i) - \psi(h_i^1, \dots, h_i^k, h_i^{k+1,k} - \Delta_i^{k+1}, \gamma_i)\bigr| \\ &\le L(k+2)^{\frac{r}{2}-1}\,\|\Delta^{k+1}\|_{n,r}\biggl(1 + 2\sum_{\ell=1}^{k}\|h^\ell\|_{n,r}^{r-1} + 2\|\gamma\|_{n,r}^{r-1} + \|h^{k+1,k}\|_{n,r}^{r-1} + \|h^{k+1,k} - \Delta^{k+1}\|_{n,r}^{r-1}\biggr) \\ &\lesssim_{k,r} L\|\Delta^{k+1}\|_{n,r}\biggl(1 + \sum_{\ell=1}^{k}\|h^\ell\|_{n,r}^{r-1} + \|h^{k+1,k}\|_{n,r}^{r-1} + \|\gamma\|_{n,r}^{r-1} + \|\Delta^{k+1}\|_{n,r}^{r-1}\biggr) \qquad (6.47)\end{aligned}$$

for each integer $n > k$. Now $\|h^{k+1,k}\|_{n,r} \overset{d}{=} \|h^{k+1}\|_{n,r}$ for each such n, and recall from $\mathcal{H}_{k+1}(a)$ that $\|\Delta^{k+1}\|_{n,r} = o_c(1)$ and from $\mathcal{H}_{k+1}(b)$ that $\|h^\ell\|_{n,r} = O_c(1)$ for $1 \le \ell \le k+1$. Thus, $T_{n2} \xrightarrow{c} 0$

by (6.47), and we conclude from (6.39) that $n^{-1}\sum_{i=1}^n \psi(h_i^1,\dots,h_i^k,h_i^{k+1},\gamma_i) \overset{d}{=} T_n = T_{n1} + T_{n2} \overset{c}{\to} \mathbb{E}\big(\psi(\bar{G}_1,\dots,\bar{G}_k,\bar{G}_{k+1},\bar{\gamma})\big)$, as required.

$\mathcal{H}_{k+1}(d)$: The arguments in this proof are similar to those given for $\mathcal{H}_1(d)$ and $\mathcal{H}_{k+1}(c)$, so we outline the key steps without going into the full details. By (6.38) and Lemma 7.6(c), we have $(\gamma,m^0,h^1,\dots,h^k,h^{k+1}) \overset{d}{=}|_{\mathcal{S}_k} (\gamma,m^0,h^1,\dots,h^k,h^{k+1,k})$ for each $n > k$, so for fixed $\phi \in \mathrm{PL}_{k+2}(1)$, it follows that $n^{-1}\sum_{i=1}^n m_i^0\,\phi(h_i^1,\dots,h_i^k,h_i^{k+1}) \overset{d}{=} n^{-1}\sum_{i=1}^n m_i^0\,\phi(h_i^1,\dots,h_i^k,h_i^{k+1,k}) =: T_n$ for each such n. Using (6.38), we now write

$$\begin{aligned} T_n &= \frac{1}{n}\sum_{i=1}^n m_i^0\,\phi(h_i^1,\dots,h_i^k,\tilde{h}_i^{k+1},\gamma_i) + \frac{1}{n}\sum_{i=1}^n m_i^0\big\{\phi(h_i^1,\dots,h_i^k,h_i^{k+1,k},\gamma_i) - \phi(h_i^1,\dots,h_i^k,\tilde{h}_i^{k+1},\gamma_i)\big\} \\ &=: T_{n1} + T_{n2} \end{aligned} \tag{6.48}$$

for each $n > k$, and aim to prove that $T_{n1} \overset{c}{\to} \mathbb{E}\big(F_0(\bar{\gamma})\cdot\phi(\bar{G}_1,\dots,\bar{G}_k,\bar{\gamma})\big)$ and $T_{n2} \overset{c}{\to} 0$, which together imply the desired conclusion.

For T_{n1}, recall once again from (6.21) or (6.38) that $\tilde{h}^{k+1}(n) = \sum_{\ell=1}^k \bar{\alpha}_\ell^k\,h^\ell(n) + \overset{\perp}{\tau}_{k+1}\tilde{Z}^{k+1}(n)$ for each n, where $\tilde{Z}^{k+1} \equiv \tilde{Z}^{k+1}(n) \sim N_n(0,I_n)$ is independent of $\mathcal{S}_k \equiv \mathcal{S}_k(n) = \sigma(\gamma,m^0,h^1,\dots,h^k)$. Define $\Phi\colon \mathbb{R}^{k+1} \to \mathbb{R}$ by $\Phi(x_1,\dots,x_k,y) := \mathbb{E}\big\{\phi\big(x_1,\dots,x_k,\sum_{\ell=1}^k \bar{\alpha}_\ell^k\,x_\ell + \overset{\perp}{\tau}_{k+1}Z,y\big)\big\}$ with $Z \sim N(0,1)$. Then $\Phi \in \mathrm{PL}_{k+1}(1)$ by Lemma 7.23(b), and as in (6.40) and (6.41), it follows from the inductive hypothesis $\mathcal{H}_k(d)$ and Lemma 7.7 that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n m_i^0\,\mathbb{E}\big(\phi(h_i^1,\dots,h_i^k,\tilde{h}_i^{k+1},\gamma_i)\,\big|\,\mathcal{S}_k\big) &= \frac{1}{n}\sum_{i=1}^n m_i^0\,\Phi(h_i^1,\dots,h_i^k,\gamma_i) \\ &\overset{c}{\to} \mathbb{E}\big(F_0(\bar{\gamma})\cdot\Phi(\bar{G}_1,\dots,\bar{G}_k,\bar{\gamma})\big) \\ &= \mathbb{E}\big(F_0(\bar{\gamma})\cdot\phi(\bar{G}_1,\dots,\bar{G}_k,\bar{G}_{k+1},\bar{\gamma})\big). \end{aligned} \tag{6.49}$$

Next, we show as in (6.42) that

$$T_{n1}' := \frac{1}{n}\sum_{i=1}^n m_i^0\big\{\phi(h_i^1,\dots,h_i^k,\tilde{h}_i^{k+1},\gamma_i) - \mathbb{E}\big(\phi(h_i^1,\dots,h_i^k,\tilde{h}_i^{k+1},\gamma_i)\,\big|\,\mathcal{S}_k\big)\big\} \overset{c}{\to} 0. \tag{6.50}$$

To this end, for each $u \in \mathbb{R}$ and $v \equiv (x_1,\dots,x_k,y) \in \mathbb{R}^{k+1}$, define $\phi_{u,v},\bar{\phi}_{u,v}\colon \mathbb{R} \to \mathbb{R}$ by $\phi_{u,v}(z) := u\phi\big(x_1,\dots,x_k,\sum_{\ell=1}^k \bar{\alpha}_\ell^k\,x_\ell + \overset{\perp}{\tau}_{k+1}z,y\big)$ and $\bar{\phi}_{u,v}(z) := \phi_{u,v}(z) - \mathbb{E}\big(\phi_{u,v}(Z)\big)$, where $Z \sim N(0,1)$. Since $\phi \in \mathrm{PL}_{k+2}(1)$, we deduce from Lemma 7.23(a) that there exists $K' > 0$, depending only on the deterministic $\bar{\alpha}^k \equiv (\bar{\alpha}_1^k,\dots,\bar{\alpha}_k^k)$, $\overset{\perp}{\tau}_{k+1}$ and r, such that $\bar{\phi}_{u,v} \in \mathrm{PL}_1(1,LK'|u|)$ for each $u \in \mathbb{R}$ and $v \in \mathbb{R}^{k+1}$. Now define the $\mathcal{S}_k$-measurable vectors $v_k^{(i)} := (h_i^1,\dots,h_i^k,\gamma_i)$ for $1 \le i \le n$, as in (6.45), and let $\tilde{E}_r$ be as in (6.34). Then by Lemma 7.7 and (7.12) in Remark 7.13, it follows as in (6.34), (6.44) and (6.45) that for each $n > k$ and $\varepsilon > 0$, we have

$$\mathbb{P}(|T_{n1}'| > \varepsilon\,|\,\mathcal{S}_k) = \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \bar{\phi}_{m_i^0,v_k^{(i)}}(\tilde{Z}_i^{k+1})\right| \ge \varepsilon \;\middle|\; \mathcal{S}_k\right) \le \tilde{E}_r(n,\varepsilon,LK'\|m^0\|_n). \tag{6.51}$$

Now $m^0 \equiv m^0(n)$ is measurable with respect to $\mathcal{S}_0 \subseteq \mathcal{S}_k \equiv \mathcal{S}_k(n)$ for each n, and $\|m^0\|_n = O_c(1)$ by (A2), so we conclude as in (6.36) and (6.37) that $\sum_n \mathbb{P}(|T_{n1}'| > \varepsilon) < \infty$ for all $\varepsilon > 0$. Thus, $T_{n1}' \overset{c}{\to} 0$ by Proposition 1.2, as claimed in (6.50).

Finally, we prove that $T_{n2} \overset{c}{\to} 0$. Let $L > 0$ be such that $\phi \in \mathrm{PL}_{k+2}(1,L)$, and for $n > k$ and $1 \le i \le n$, define $v_{k+1}^{(i)} = (h_i^1,\dots,h_i^k,h_i^{k+1,k},\gamma_i)$ and $\tilde{v}_{k+1}^{(i)} = (h_i^1,\dots,h_i^k,\tilde{h}_i^{k+1},\gamma_i) = (h_i^1,\dots,h_i^k,h_i^{k+1,k} - \Delta_i^{k+1},\gamma_i)$ as in (6.47), where the final equality is obtained from (6.22). As in the proof of $\mathcal{H}_1(d)$, we now apply the Cauchy–Schwarz inequality and the fact that $\|\cdot\|_n \equiv \|\cdot\|_{n,2} \le \|\cdot\|_{n,r}$ to see that

$$|T_{n2}| \le \frac{1}{n}\sum_{i=1}^n |m_i^0|\big|\phi(v_{k+1}^{(i)}) - \phi(\tilde{v}_{k+1}^{(i)})\big| \le \frac{L}{n}\sum_{i=1}^n |m_i^0||\Delta_i^{k+1}| \le L\|m^0\|_n\|\Delta^{k+1}\|_n.$$

Since $\|m^0\|_n \xrightarrow{c} \tau_1$ by (A2) and $\|\Delta^{k+1}\|_n \leq \|\Delta^{k+1}\|_{n,r} = o_c(1)$ by $\mathcal{H}_{k+1}(a)$, we conclude that $T_{n2} = o_c(1)$, as required. Together with (6.48), (6.49), (6.50), this yields $\mathcal{H}_{k+1}(d)$, and hence completes the inductive step for Proposition 6.16. □

Proof of Remark 6.17. Under (A0), (A4) and (A5), if instead (A1)–(A3) hold with $\xrightarrow{c}$ and $O_c(1)$ replaced with $\xrightarrow{p}$ and $O_p(1)$ respectively, then as explained in the third bullet point in Remark 6.1, we can make the same replacements in the proof of Proposition 6.16 and most of the arguments go through as before. However, a few alterations are required in the proofs of (c, d) and (g), which we now describe.

First, in the proof of $\mathcal{H}_1(d)$, the goal in (6.33) is now to show that $T'_{n1} \xrightarrow{p} 0$ as $n \to \infty$. Instead of proceeding as in (6.36) and (6.37), we return to (6.35), where we note that if $\varepsilon > 0$ is fixed and (A2) takes the form $\|m^0\|_n \xrightarrow{p} \tau_1$, then $\mathbb{P}(|T'_{n1}| > \varepsilon \,|\, \mathcal{S}_0) \leq \tilde{E}_r(n, \varepsilon, L\tau_1\|m^0\|_n) = o_p(1)$ by Slutsky's lemma and the definition of $\tilde{E}_r$ in (6.34). Then for every $\varepsilon > 0$, it follows from the bounded convergence theorem that $\mathbb{P}(|T'_{n1}| > \varepsilon) = \mathbb{E}\bigl(\mathbb{P}(|T'_{n1}| > \varepsilon \,|\, \mathcal{S}_0)\bigr) \to 0$ as $n \to \infty$, so $T'_{n1} \xrightarrow{p} 0$, as desired.

In the proofs of $\mathcal{H}_{k+1}(c, d)$, the analogues of (6.42) and (6.50) can be derived from (6.45, 6.46) and (6.51) respectively in much the same way; for the former, since $\|h^\ell\|_{n,r} = O_p(1)$ for $1 \leq \ell \leq k$ by the modified $\mathcal{H}_k(b)$, (6.46) implies that $\check{L}_k(n) \lesssim_{k,r} LK\bigl(1 + \sum_{\ell=1}^k \|h^\ell\|_{n,r}^r\bigr)^{1/r'} = O_p(1)$.

In addition, $\mathcal{H}_k(g)$ now reads $b_k = \langle f'_k(h^k, \gamma)\rangle_n \xrightarrow{p} \mathbb{E}\bigl(f'_k(\bar{G}_k, \bar{\gamma})\bigr) = \bar{b}_k$. To prove this, we can argue along subsequences, similarly to the proof of Corollary 7.21(b). □

Proofs of Theorems 2.1 and 2.3. Theorem 2.1 follows from Theorem 2.3, which in turn is a immediate consequence of Proposition 6.16(c) and Corollary 7.21(b). □

Proofs for Remark 6.1. (a) *Convergence in probability*: This is immediate from Remark 6.17 and Corollary 7.21(b).

(b) *Almost sure convergence*: The random sequences $\Upsilon := \bigl(m^0(n) : n \in \mathbb{N}\bigr)$ and $\Gamma := \bigl(\gamma(n) : n \in \mathbb{N}\bigr)$ take values in $E := \prod_{n=1}^\infty \mathbb{R}^n$, whose cylindrical and Borel σ-algebras coincide by Kallenberg (1997, Lemma 1.2). Let E^* be the set of all $(u, v) \in E \times E$ such that (A1)–(A3) hold when $\Upsilon = u \equiv \bigl(u(n) : n \in \mathbb{N}\bigr)$ and $\Gamma = v \equiv \bigl(v(n) : n \in \mathbb{N}\bigr)$ are non-random. It can be verified that E^* is a Borel subset of $E \times E$.

For $k \in \mathbb{N}$, let $\bar{\mu}^k := N(0, \tau_k^2) \otimes \pi$ and $\mu_n^k := \nu_n(h^k, \gamma)$ for $n \in \mathbb{N}$. In the special case where $(\Upsilon, \Gamma) \in E^*$ is deterministic, Theorem 2.1 implies that for each k, the resulting sequence of AMP iterates $\bigl(h^k(n) : n \in \mathbb{N}\bigr)$ satisfies $d_r(\mu_n^k, \bar{\mu}^k) \xrightarrow{c} 0$. Note that for each n, we can write $d_r(\mu_n^k, \bar{\mu}^k) = d_r\bigl(\nu_n(h^k, \gamma), \bar{\mu}^k\bigr) = g_n\bigl(m^0(n), \gamma(n), W(n)\bigr)$ for some (non-random) Borel measurable $g_n \colon \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{n \times n} \to \mathbb{R}$. Indeed, we see from (2.1) that $h^k \equiv h^k(n)$ is a deterministic Borel measurable function of $m^0(n)$, $\gamma(n)$ and $W(n)$. Moreover for all $x, x' \in \mathbb{R}^n$ and the corresponding empirical distributions $\nu_n(x), \nu_n(x')$ of their components, we have

$$\bigl|d_r\bigl(\nu_n(x), \bar{\mu}^k\bigr) - d_r\bigl(\nu_n(x'), \bar{\mu}^k\bigr)\bigr| \leq d_r\bigl(\nu_n(x), \nu_n(x')\bigr) \leq \bigl(n^{-1} \textstyle\sum_{i=1}^n |x_i - x'_i|^r\bigr)^{1/r} = \|x - x'\|_{n,r}$$

since d_r is a metric, so $x \mapsto d_r\bigl(\nu_n(x), \bar{\mu}^k\bigr)$ is continuous on $\mathbb{R}^n$.

Since $\{(a_1, a_2, \dots) \in \mathbb{R}^{\mathbb{N}} : \lim_{n\to\infty} a_n = 0\}$ is a Borel subset of $\mathbb{R}^{\mathbb{N}}$, we conclude that $g \colon (u, v) \mapsto \mathbb{P}\bigl\{\lim_{n\to\infty} g_n\bigl(u(n), v(n), W(n)\bigr) = 0\bigr\}$ is a well-defined Borel measurable function on $E \times E$ satisfying $g(u, v) = 1$ for all $(u, v) \in E^*$.

Now suppose more generally that (Υ, Γ) and $\bigl(W(n) : n \in \mathbb{N}\bigr)$ are independent. If $(\Upsilon, \Gamma) \in E^*$ almost surely, then for the corresponding sequence of AMP iterates $\bigl(h^k(n) : n \in \mathbb{N}\bigr)$ from (2.1),

$$\begin{aligned}
\mathbb{P}\Bigl(\lim_{n\to\infty} d_r(\mu_n^k, \bar{\mu}^k) = 0\Bigr) &= \mathbb{P}\Bigl(\lim_{n\to\infty} g_n\bigl(m^0(n), \gamma(n), W(n)\bigr) = 0\Bigr) \\
&= \mathbb{E}\Bigl\{\mathbb{P}\Bigl(\lim_{n\to\infty} g_n\bigl(m^0(n), \gamma(n), W(n)\bigr) = 0 \Bigm| \Upsilon, \Gamma\Bigr)\Bigr\} \\
&= \mathbb{E}\bigl(g(\Upsilon, \Gamma)\bigr) \geq \mathbb{E}\bigl(g(\Upsilon, \Gamma)\,\mathbb{1}_{\{(\Upsilon,\Gamma) \in E^*\}}\bigr) = 1,
\end{aligned}$$

where the third equality follows from the independence assumption and Lemma 7.7. Therefore, $d_r(\mu_n^k, \bar{\mu}^k) \overset{a.s.}{\to} 0$ as $n \to \infty$, and moreover $\widetilde{d}_r(\mu_n^k, \bar{\mu}^k) \overset{a.s.}{\to} 0$ by Corollary 7.21(a). □

6.6 Auxiliary results and proofs for Section 2

Proof of Lemma 2.2. We proceed by induction on k, noting first that the base case $k = 1$ is trivial since $\bar{\mathrm{T}}^{[1]} = \tau_1^2 > 0$ by (A4). Now for $k \geq 2$ and $a \equiv (a_1, \dots, a_k) \in \mathbb{R}^k$, recall the expression (2.7) for $a^\top \bar{\mathrm{T}}^{[k]} a$. If $a_1 \neq 0 = a_2 = \cdots = a_k$, then $a^\top \bar{\mathrm{T}}^{[k]} a = (a_1\tau_1)^2 > 0$ as in the base case. On the other hand, if $a_L \neq 0$ for some $2 \leq L \leq k$, then by (A4), we can find $B_L \subseteq \mathbb{R}$ with $\pi(B_L) > 0$ such that $x_{L-1} \mapsto a_L f_{L-1}(x_{L-1}, y)$ is non-constant whenever $y \in B_L$. For all such y, note that $(x_1, \dots, x_{k-1}) \mapsto \sum_{\ell=2}^k a_\ell f_{\ell-1}(x_{\ell-1}, y)$ is non-constant on $\mathbb{R}^{k-1}$. Now by the inductive hypothesis, $(G_1, \dots, G_{k-1})$ has a positive definite covariance matrix $\bar{\mathrm{T}}^{[k-1]}$, so the random variable $a_1 F_0(y) + \sum_{\ell=2}^k a_\ell f_{\ell-1}(G_{\ell-1}, y)$ is non-degenerate whenever $y \in B_L$. Since $\bar{\gamma} \sim \pi$ is independent of $G_1, \dots, G_{k-1}$ and $\mathbb{P}(\bar{\gamma} \in B_L) = \pi(B_L) > 0$, it follows that $a_1 F_0(\bar{\gamma}) + \sum_{\ell=2}^k a_\ell f_{\ell-1}(G_{\ell-1}, \bar{\gamma})$ is also non-degenerate. Thus, in all cases, it follows from (2.7) and (A3) that $a^\top \bar{\mathrm{T}}^{[k]} a > 0$ whenever $a \neq 0$, as claimed. □

Proof of Remark 6.5. Since $\tilde{f}_0$ is Lipschitz and $\bar{\eta}, \bar{\gamma} \in \mathcal{P}_1(2)$, we have $\mathbb{E}\{\|(\tilde{f}_0(\bar{\eta}, \bar{\gamma}), \bar{\gamma})\|^2\} < \infty$, so $\mu^0 \in \mathcal{P}_2(2)$. By Corollary 7.21(b), an equivalent formulation of (A1$^+$) is that

$$\frac{1}{n}\sum_{i=1}^n \psi(m_i^0, \gamma_i) \xrightarrow{c} \mathbb{E}\{\psi(\tilde{f}_0(\bar{\eta}, \bar{\gamma}), \bar{\gamma})\}$$

for all $\psi \in \mathrm{PL}_2(2)$. In particular, $(x, y) \mapsto x^2$ lies in $\mathrm{PL}_2(2)$, so $\|m^0\|_n^2 = n^{-1}\sum_{i=1}^n |m_i^0|^2 \xrightarrow{c} \mathbb{E}(\tilde{f}_0(\bar{\eta}, \bar{\gamma})^2) =: \tau_1^2$, which yields the first part of (A2). Moreover, the function $F_0 \colon \mathbb{R} \to \mathbb{R}$ defined by $F_0(y) := \mathbb{E}(\tilde{f}_0(\bar{\eta}, y))$ is Lipschitz, and since $\bar{\eta}, \bar{\gamma}$ are independent, it follows from Lemma 7.7 and Jensen's inequality that

$$\mathbb{E}(F_0(\bar{\gamma})^2) = \mathbb{E}\{\mathbb{E}(\tilde{f}_0(\bar{\eta}, \bar{\gamma}) \,|\, \bar{\gamma})^2\} \leq \mathbb{E}\{\mathbb{E}(\tilde{f}_0(\bar{\eta}, \bar{\gamma})^2 \,|\, \bar{\gamma})\} = \tau_1^2.$$

For each Lipschitz $\phi \colon \mathbb{R} \to \mathbb{R}$, Lemma 7.22 ensures that $(x, y) \mapsto x\phi(y)$ belongs to $\mathrm{PL}_2(2)$, so

$$\langle m^0, \phi(\gamma)\rangle_n = \frac{1}{n}\sum_{i=1}^n m_i^0\, \phi(\gamma_i) \xrightarrow{c} \mathbb{E}(\tilde{f}_0(\bar{\eta}, \bar{\gamma}) \cdot \phi(\bar{\gamma})) = \mathbb{E}\{\mathbb{E}(\tilde{f}_0(\bar{\eta}, \bar{\gamma}) \,|\, \bar{\gamma})\phi(\bar{\gamma})\} = \mathbb{E}(F_0(\bar{\gamma})\phi(\bar{\gamma})),$$

where the final equality again follows from Lemma 7.7. Therefore, (A3) also holds. □

The following auxiliary result is used in the proof of Proposition 6.16(a) to control the third summand in the deviation term Δ^{k+1} defined in (6.22).

Lemma 6.18. *For $n \in \mathbb{N}$ and $k \in \{0, 1 \dots, n-1\}$, let $\tilde{Z}^{k+1} \equiv \tilde{Z}^{k+1}(n)$ be as in Proposition 6.11, so that $\tilde{Z}^{k+1} \sim N_n(0, I_n)$ is independent of $\mathcal{S}_k$. If $\xi_1, \dots, \xi_k \overset{\mathrm{iid}}{\sim} N(0,1)$ and $r \geq 1$, then $\|P_k \tilde{Z}^{k+1}\|_{n,r}$ is stochastically dominated by $\sum_{i=1}^k |\xi_i|/n^{1/(r\vee 2)}$.*

Proof. Note that P_k is an $\mathcal{S}_k$-measurable projection matrix of rank $r_k \leq k$. Since $\tilde{Z}^{k+1}$ is independent of $\mathcal{S}_k$, it therefore has conditional distribution $N_n(0, I_n)$ given $\mathcal{S}_k$ by Remark 7.4. Now let $\xi_1, \dots, \xi_k \overset{\mathrm{iid}}{\sim} N(0,1)$ be independent of $\mathcal{S}_k$ and let $\{\tilde{m}^1, \dots, \tilde{m}^{r_k}\}$ be any $\mathcal{S}_k$-measurable orthonormal basis of $\mathrm{Im}(M_k) = \mathrm{Im}(P_k)$, as in Remark 6.10. Recall that if $Z \sim N_n(0, I_n)$ and $P \in \mathbb{R}^{n \times n}$ is a deterministic projection matrix of rank p, then $PZ \sim N(0, P)$, which is also the distribution of $\sum_{i=1}^p \zeta_i u_i$ when $\zeta_1, \dots, \zeta_p \overset{\mathrm{iid}}{\sim} N(0,1)$ and $\{u_1, \dots, u_p\}$ is any orthonormal basis of $\mathrm{Im}(P)$. We deduce from this and Lemma 7.6(b) that $P_k \tilde{Z}^{k+1}$ and $\sum_{i=1}^{r_k} \xi_i \tilde{m}^i$ both have conditional distribution $N_n(0, P_k)$ given $\mathcal{S}_k$. This implies that $\|P_k \tilde{Z}^{k+1}\|_{n,r} \overset{d}{=} \|\sum_{i=1}^{r_k} \xi_i \tilde{m}^i\|_{n,r}$.

Now for all $x \in \mathbb{R}^n$, we have $\|x\|_{n,r} = n^{-1/r}\|x\|_r \leq n^{-1/(r\vee 2)}\|x\|_2$ by Hölder's inequality and the fact that $\|\cdot\|_{p'} \leq \|\cdot\|_p$ for $1 \leq p \leq p'$. Since $\|\tilde{m}^i\|_2 = 1$ for all i by definition, it follows from this and the triangle inequality for $\|\cdot\|_{n,r}$ that $\|\sum_{i=1}^{r_k} \xi_i \tilde{m}^i\|_{n,r} \leq \sum_{i=1}^{r_k} |\xi_i|\, \|\tilde{m}^i\|_{n,r} \leq \sum_{i=1}^{r_k} |\xi_i|/n^{1/(r\vee 2)} \leq \sum_{i=1}^k |\xi_i|/n^{1/(r\vee 2)}$. Combining this with the conclusion of the previous paragraph yields the result. □

6.7 AMP with matrix-valued iterates

As mentioned in Section 2.1, state evolution characterisations can be obtained for more general abstract AMP recursions in which the iterates are matrices rather than vectors. Here, we will briefly describe the extended version of the asymmetric iteration (2.10), which is used to establish the master theorem for GAMP in Section 4.1.

For $n, p \in \mathbb{N}$, let $W \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}^n$ be as in (B0). For $\ell_E, \ell_H \in \mathbb{N}$, let $(g_k, f_{k+1} : k \in \mathbb{N}_0)$ be two sequences of Lipschitz functions $g_k \colon \mathbb{R}^{\ell_E} \times \mathbb{R} \to \mathbb{R}^{\ell_H}$ and $f_{k+1} \colon \mathbb{R}^{\ell_H} \times \mathbb{R} \to \mathbb{R}^{\ell_E}$, which are applied row-wise to matrices. Given $Q^{-1} := 0 \in \mathbb{R}^{n\times \ell_H}$, $B_0 \in \mathbb{R}^{\ell_E\times\ell_H}$ and $M^0 \in \mathbb{R}^{p\times\ell_E}$, inductively define

$$\begin{aligned} E^k &:= WM^k - Q^{k-1}B_k^\top, & Q^k &:= g_k(E^k, \gamma), & C_k &:= n^{-1}\textstyle\sum_{i=1}^n g_k'(E_i^k, \gamma_i), \\ H^{k+1} &:= W^\top Q^k - M^k C_k^\top, & M^{k+1} &:= f_{k+1}(H^{k+1}, \beta), & B_{k+1} &:= n^{-1}\textstyle\sum_{j=1}^p f_{k+1}'(H_j^{k+1}, \beta_j) \end{aligned} \tag{6.52}$$

for $k \in \mathbb{N}_0$. Here, E_i^k and H_j^{k+1} denote the i^{th} and j^{th} rows of $E^k \in \mathbb{R}^{n\times\ell_E}$ and $H^{k+1} \in \mathbb{R}^{p\times\ell_H}$ respectively. Also, $g_k' \colon \mathbb{R}^{\ell_E}\times\mathbb{R} \to \mathbb{R}^{\ell_H\times\ell_E}$ and $f_{k+1}' \colon \mathbb{R}^{\ell_H}\times\mathbb{R} \to \mathbb{R}^{\ell_E\times\ell_H}$ are bounded, Borel measurable functions that agree with the derivatives (Jacobians) of g_k, f_{k+1} respectively with respect to their first arguments, wherever the latter are defined.

Consider now a sequence of recursions (6.52) indexed by n and $p \equiv p_n$ with $n/p \to \delta \in (0,\infty)$ as $n \to \infty$, and assume appropriate analogues of (B0)–(B5) with $r \in [2,\infty)$. In particular, suppose in place of (B2) that $(M^0)^\top M^0/n \overset{c}{\to} \Sigma_0$ for some non-negative definite $\Sigma_0 \in \mathbb{R}^{\ell_E\times\ell_E}$, and that $p^{-1}\sum_{i=1}^p\sum_{j=1}^{\ell_E}|M_{ij}^0|^r = O_c(1)$. The state evolution recursion for (6.52) is then defined analogously to that in (2.11), via

$$\begin{aligned} \mathrm{T}_{k+1} &:= \mathbb{E}\big(g_k(G_k^\sigma, \bar\gamma)\, g_k(G_k^\sigma, \bar\gamma)^\top\big) \in \mathbb{R}^{\ell_H\times\ell_H}, \\ \Sigma_{k+1} &:= \delta^{-1}\,\mathbb{E}\big(f_{k+1}(G_{k+1}^\tau, \bar\beta) f_{k+1}(G_{k+1}^\tau, \bar\beta)^\top\big) \in \mathbb{R}^{\ell_E\times\ell_E} \end{aligned} \tag{6.53}$$

for $k \in \mathbb{N}_0$, where we take $G_k^\sigma \sim N_{\ell_E}(0, \Sigma_k)$ to be independent of $\bar\gamma \sim \pi_{\bar\gamma}$, and $G_{k+1}^\tau \sim N_{\ell_H}(0, \mathrm{T}_{k+1})$ to be independent of $\bar\beta \sim \pi_{\bar\beta}$.

For $k \in \mathbb{N}_0$, it can be shown that the empirical distributions of the rows of $(E^k\ \gamma)$ and $(H^{k+1}\ \beta)$ converge completely in d_r to $N_{\ell_E}(0,\Sigma_k)\otimes\pi_\gamma$ and $N_{\ell_H}(0,\mathrm{T}_{k+1})\otimes\pi_\beta$ respectively as $n, p \to \infty$ with $n/p \to \delta$. Similarly as in Remark 2.4, these limiting distributions remain unchanged if one or both of C_k, B_{k+1} are replaced with the deterministic matrices $\bar C_k := \mathbb{E}\big(g_k'(G_k^\Sigma, \bar\gamma)\big)$ and $\bar B_{k+1} := \delta^{-1}\mathbb{E}\big(f_{k+1}'(G_{k+1}^\mathrm{T}, \bar\beta)\big)$ respectively. Moreover, by generalising the definitions (2.12)–(2.13) of the limiting covariance matrices in line with (6.53), one can obtain the d_r limits of the joint empirical distributions for (6.52) above.

The proofs of these results are conceptually very similar to that of Theorem 2.5. For further details, see Javanmard and Montanari (2013), who first consider a generalisation of the symmetric iteration (2.1) with matrix-valued iterates, and then handle the asymmetric case by a reduction argument.

6.8 Proofs for Section 3

Proof of Theorem 3.1. As described in the proof sketch on page 16, we introduce the recursion (3.13) given by $u^1 \equiv u^1(n) = W\hat v^0 = Wg_0(v^0)$ and

$$\begin{aligned} u^{k+1} \equiv u^{k+1}(n) &= Wg_k(u^k + \mu_k v) - \tilde b_k g_{k-1}(u^{k-1} + \mu_{k-1}v) \\ &= Wf_k(u^k, v) - \tilde b_k f_{k-1}(u^{k-1}, v) \end{aligned}$$

for $k, n \in \mathbb{N}$, where $f_k(x,y) = g_k(x + \mu_k y)$ and $f_k'(x,y) = g_k'(x + \mu_k y)$ for $x, y \in \mathbb{R}$, and $\tilde b_k \equiv \tilde b_k(n) = \langle g_k'(u^k + \mu_k v)\rangle_n = \langle f_k'(u^k, v)\rangle_n$. First, we verify that this is an iteration of the form (2.1) to which we can apply the master theorems from Section 2.1 for symmetric AMP. Indeed, it follows from (M0) and (M1) respectively that (3.13) satisfies (A0) and (A1$^+$), where the latter holds with $m^0 = v$, $\gamma = v$, $\bar{\gamma} = V \sim \pi$, $\bar{\eta} = U$ and $\tilde{f}_0(x,y) = f_0(\mu_0 x + \sigma_0 y)$ for $x, y \in \mathbb{R}$. By Remark 6.5, (A1$^+$) implies that (A1)–(A3) hold with $r = 2$ and $\tau_1 = \text{c-lim}_{n\to\infty} \|\hat{v}^0\|_n^2 = \sigma_1^2$. As verified in (3.14), the state evolution parameters $(\tau_k : k \in \mathbb{N})$ for (3.13) satisfy $\tau_k^2 = \sigma_k^2$ for all k in view of (3.6). Finally, by (M2), each $f_k \colon \mathbb{R}^2 \to \mathbb{R}$ is Lipschitz and the corresponding f_k' satisfies (A5).

Consequently, for each $k \in \mathbb{N}$, it follows from Theorem 2.3 that

$$\sup_{\psi \in \mathrm{PL}_{k+1}(2,1)} \left| \frac{1}{n} \sum_{i=1}^n \psi(v_i^0, v_i^1, \dots, v_i^k, v_i) - \mathbb{E}\big(\psi(\mu_0 V + \sigma_0 U, \sigma_1 G_1, \dots, \sigma_k G_k, V)\big) \right| \xrightarrow{c} 0$$

as $n \to \infty$, where $(\sigma_1 G_1, \dots, \sigma_k G_k) \sim N_k(0, \bar{\Sigma}^{[k]})$ is taken to be independent of (U, V) from (M1). Since $\Phi_k \colon (x_1, \dots, x_k, y) \mapsto (x_1 + \mu_1 y, \dots, x_k + \mu_k y, y)$ is a linear map with Lipschitz constant $\tilde{L}_k := \|(\mu_1, \dots, \mu_k, 1)\|$, we have $\tilde{L}_k^{-2}(\psi \circ \Phi_k) \in \mathrm{PL}_{k+1}(2,1)$ whenever $\psi \in \mathrm{PL}_{k+1}(2,1)$, so it follows from the display above that

$$\begin{aligned} \sup_{\psi \in \mathrm{PL}_{k+2}(2,1)} \bigg| \frac{1}{n} \sum_{i=1}^n &\psi(v_i^0, v_i^1 + \mu_1 v_i, \dots, v_i^k + \mu_k v_i, v_i) \\ &- \mathbb{E}\big(\psi(\mu_0 V + \sigma_0 U, \mu_1 V + \sigma_1 G_1, \dots, \mu_k V + \sigma_k G_k, V)\big) \bigg| \xrightarrow{c} 0 \end{aligned} \tag{6.54}$$

as $n \to \infty$. Defining $\tilde{\Delta}^k \equiv \tilde{\Delta}^k(n) := v^k - (u^k + \mu_k v) \in \mathbb{R}^n$ for $k, n \in \mathbb{N}$, we can apply Lemma 7.24 to see that

$$\begin{aligned} \sup_{\psi \in \mathrm{PL}_{k+2}(2,1)} &\left| \frac{1}{n} \sum_{i=1}^n \psi(v_i^0, v_i^1, \dots, v_i^k, v_i) - \psi(v_i^0, v_i^1 + \mu_1 v_i, \dots, v_i^k + \mu_k v_i, v_i) \right| \\ &\le \left(\sum_{\ell=1}^k \|\tilde{\Delta}^\ell\|_n^2 \right)^{1/2} \left(1 + \sum_{\ell=1}^k \big(\|v^\ell\|_n + \|u^\ell + \mu_\ell v\|_n\big) + 2\big(\|v^0\|_n + \|v\|_n\big) \right) \\ &\le \left(\sum_{\ell=1}^k \|\tilde{\Delta}^\ell\|_n^2 \right)^{1/2} \left(1 + \sum_{\ell=1}^k \big(\|\tilde{\Delta}^\ell\|_n + 2\|u^\ell + \mu_\ell v\|_n\big) + 2\big(\|v^0\|_n + \|v\|_n\big) \right) \end{aligned} \tag{6.55}$$

for all k and n, where $\|\cdot\|_n \equiv \|\cdot\|_{n,2} = n^{-1/2}\,\|\cdot\|$ on $\mathbb{R}^n$. For every $\ell \in \mathbb{N}$, it follows from (M1) and (6.54) that

$$\begin{aligned} \|v^0\|_n &\xrightarrow{c} \mathbb{E}\big((\mu_0 V + \sigma_0 U)^2\big) = \mu_0^2 + \sigma_0^2, \qquad \|v\|_n^2 \to \mathbb{E}(V^2) = 1 \\ \text{and} \quad \|u^\ell + \mu_\ell v\|_n^2 &= \frac{1}{n} \sum_{i=1}^n (u_i^\ell + \mu_\ell v_i)^2 \xrightarrow{c} \mathbb{E}\big((\mu_\ell V + \sigma_\ell G_\ell)^2\big) = \mu_\ell^2 + \sigma_\ell^2 \end{aligned} \tag{6.56}$$

as $n \to \infty$. We will now establish by induction on $k \in \mathbb{N}$ that

$$\|\tilde{\Delta}^k\|_n = \|v^k - (u^k + \mu_k v)\|_n \xrightarrow{c} 0 \quad \text{as } n \to \infty \tag{6.57}$$

and hence that the conclusion (3.8) of Theorem 3.1 holds for every k. For the base case $k = 1$, we have $\|v\|_n \xrightarrow{c} 1$ by (3.4) and $\lambda\langle \hat{v}^0, v\rangle_n \xrightarrow{c} \mu_1$ by (3.5), so

$$\|\tilde{\Delta}^1\|_n = \|v^1 - (u^1 + \mu_1 v)\|_n = \|A\hat{v}^0 - (W\hat{v}^0 + \mu_1 v)\|_n = |\lambda\langle \hat{v}^0, v\rangle_n - \mu_1|\,\|v\|_n \xrightarrow{c} 0$$

as $n \to \infty$. It follows from this and (6.54)–(6.56) that (3.8) holds when $k = 1$. For a general $k \ge 2$, we write

$$\begin{aligned} \tilde{\Delta}^{k+1} \equiv \tilde{\Delta}^{k+1}(n) &= v^{k+1} - (u^{k+1} + \mu_{k+1} v) \\ &= A g_k(v^k) - b_k g_{k-1}(v^{k-1}) - \big(W g_k(u^k + \mu_k v) - \tilde{b}_k g_{k-1}(u^{k-1} + \mu_{k-1} v) + \mu_{k+1} v\big) \\ &= \big(\lambda\langle v, g_k(v^k)\rangle_n - \mu_{k+1}\big) v + W\big(g_k(v^k) - g_k(u^k + \mu_k v)\big) \\ &\qquad\qquad + \big(\tilde{b}_k g_{k-1}(u^{k-1} + \mu_{k-1} v) - b_k g_{k-1}(v^{k-1})\big) \\ &=: R_{n1} + R_{n2} + R_{n3} \end{aligned} \tag{6.58}$$

for each $n \in \mathbb{N}$, and consider R_{n1}, R_{n2}, R_{n3} in turn. First, since $(x_1, \ldots, x_k, y) \mapsto y g_k(x_k)$ belongs to $\mathrm{PL}_{k+1}(2)$ in view of Lemma 7.22, it follows from the inductive hypothesis (3.8) and the definition of μ_{k+1} in (3.6) that $\lambda\langle v, g_k(v^k)\rangle_n = n^{-1}\sum_{i=1}^n \lambda v_i\, g_k(v_i^k) \xrightarrow{c} \lambda\mathbb{E}\bigl(V g_k(\mu_k V + \sigma_k G_k)\bigr) = \mu_{k+1}$. Together with (6.56), this implies that $\|R_{n1}\|_n = |\lambda\langle v, g_k(v^k)\rangle_n - \mu_{k+1}|\,\|v\|_n \xrightarrow{c} 0$ as $n \to \infty$.

Next, since $\|W\| \equiv \|W\|_{2\to 2} = O_c(1)$ (e.g. Anderson et al., 2010; Knowles and Yin, 2013) and g_k is L_k-Lipschitz for some $L_k > 0$, the inductive hypothesis (6.57) ensures that

$$\|R_{n2}\|_n \le \|W\|\,\|g_k(v^k) - g_k(u^k + \mu_k v)\|_n \le L_k\|W\|\,\|v^k - (u^k + \mu_k v)\|_n = L_k\|W\|\,\|\tilde{\Delta}^k\|_n \xrightarrow{c} 0$$

as $n \to \infty$. Similarly, $\|\tilde{\Delta}^{k-1}\|_n \xrightarrow{c} 0$ by induction and g_{k-1} is L_{k-1}-Lipschitz for some $L_{k-1} > 0$, so as a first step towards controlling $\|R_{n3}\|_n$, we have

$$\|g_{k-1}(u^{k-1} + \mu_{k-1}v) - g_{k-1}(v^{k-1})\|_n \le L_{k-1}\|\tilde{\Delta}^{k-1}\|_n \xrightarrow{c} 0.$$

Note that since $(x_1, \ldots, x_k, y) \mapsto g_{k-1}(x_{k-1})^2$ lies in $\mathrm{PL}_{k+1}(2)$ by Lemma 7.22, it follows from (6.54) that

$$\|g_{k-1}(u^{k-1} + \mu_{k-1}v)\|_n^2 \xrightarrow{c} \mathbb{E}\bigl(g_{k-1}(\mu_{k-1}V + \sigma_{k-1}G_{k-1})^2\bigr) = \sigma_k^2$$

as $n \to \infty$. Furthermore, since g_k' satisfies (M2), we can apply the inductive hypothesis (3.8) and argue as in the proof of Proposition 6.16(f) to see that $b_k \equiv b_k(n) = \langle g_k'(v^k)\rangle_n = n^{-1}\sum_{i=1}^n g_k'(v_i^k) \xrightarrow{c} \mathbb{E}\bigl(g_k'(\mu_k V + \sigma_k G_k)\bigr)$. Similar reasoning based on (6.54) yields $\tilde{b}_k \equiv \tilde{b}_k(n) = n^{-1}\sum_{i=1}^n g_k'(v_i^k + \mu_k v_i) \xrightarrow{c} \mathbb{E}\bigl(g_k'(\mu_k V + \sigma_k G_k)\bigr)$, so $\tilde{b}_k(n) - b_k(n) \xrightarrow{c} 0$ as $n \to \infty$. Putting everything together, we conclude that

$$\|R_{n3}\|_n \le |\tilde{b}_k - b_k|\|g_{k-1}(u^{k-1} + \mu_{k-1}v)\|_n + |b_k|\|g_{k-1}(u^{k-1} + \mu_{k-1}v) - g_{k-1}(v^{k-1})\|_n \xrightarrow{c} 0,$$

and hence that $\|\tilde{\Delta}^{k+1}\|_n \le \|R_{n1}\|_n + \|R_{n2}\|_n + \|R_{n3}\|_n \xrightarrow{c} 0$ as $n \to \infty$. Combining this with (6.54)–(6.56) yields the desired conclusion (3.8), so the inductive step is complete. □

Proof of Corollary 3.2. For $\psi \in \mathrm{PL}_2(2)$, note that since $g_k\colon \mathbb{R} \to \mathbb{R}$ is Lipschitz by assumption, $(x_0, x_1, \ldots, x_k, y) \mapsto \psi\bigl(g_k(x_k), y\bigr)$ is a $\mathrm{PL}_{k+2}(2)$ function to which we can apply (3.8). This yields (3.9), which specialises to (3.10) when $\psi = \psi_2\colon (x, y) \mapsto (x-y)^2$ is squared error loss. Finally, by considering the $\mathrm{PL}_2(2)$ functions $(x, y) \mapsto y g_k(x)$, $(x, y) \mapsto g_k(x)^2$ and $(x, y) \mapsto y^2$, we deduce from (3.9) that as $n \to \infty$, we have

$$\langle \hat{v}^k, v\rangle_n \xrightarrow{c} \mathbb{E}\bigl(V g_k(\mu_k V + \sigma_k G_k)\bigr) = \mu_{k+1}$$

as in the paragraph after (6.58) above,

$$\|\hat{v}^k\|_n^2 = \frac{1}{n}\sum_{i=1}^n g_k(v_i^k)^2 \xrightarrow{c} \mathbb{E}\bigl(g_k(\mu_k V + \sigma_k G_k)^2\bigr) = \sigma_{k+1}^2$$

and $\|v\|_n^2 \xrightarrow{c} \mathbb{E}(V^2) = 1$ as in (3.4). Combining these, we obtain (3.11). □

Proof of Lemma 3.8. Fix $\mu \ne 0$ and $\sigma > 0$, and let $V \sim \pi$ and $G \sim N(0,1)$ be independent. Then $\mu V + \sigma G$ has Lebesgue density $y \mapsto p(y) := \int_{\mathbb{R}} \phi_\sigma(y - \mu x)\, d\pi(x) > 0$, where $\phi_\sigma\colon z \mapsto (\sqrt{2\pi}\sigma)^{-1} e^{-z^2/(2\sigma^2)}$ is the density of $\sigma G \sim N(0, \sigma^2)$. Moreover, since all the derivatives of ϕ_σ are bounded on $\mathbb{R}$, we can differentiate repeatedly under the integral sign to see that $p^{(j)}(y) = \int_{\mathbb{R}} \phi_\sigma^{(j)}(y - \mu x)\, d\pi(x)$ for all y and $j \in \mathbb{N}_0$, so p is a smooth function on $\mathbb{R}$. For each $y \in \mathbb{R}$, define π_y to be the distribution on $\mathbb{R}$ with density (i.e. Radon–Nikodym derivative)

$$\frac{d\pi_y}{d\pi}\colon x \mapsto \frac{\phi_\sigma(y - \mu x)}{\int_{\mathbb{R}} \phi_\sigma(y - \mu x')\, d\pi(x')} = \frac{\phi_\sigma(y - \mu x)}{p(y)} \tag{6.59}$$

with respect to π. It is easily verified that π_y is the "conditional distribution of V given $\mu V + \sigma G = y$", formally in the sense of Remark 7.5(II). It follows from this and Dudley (2002, Theorem 10.2.5) that taking $V_y \sim \pi_y$ and defining $g(y) := \mathbb{E}(V_y) = \int_{\mathbb{R}} x\,d\pi_y(x)$ for $y \in \mathbb{R}$, we have $\mathbb{E}(V \mid \mu V + \sigma G) = g(\mu V + \sigma G)$. For each y, note that

$$g(y) = \frac{\int_{\mathbb{R}} x\,\phi_\sigma(y - \mu x)\,d\pi(x)}{\int_{\mathbb{R}} \phi_\sigma(y - \mu x)\,d\pi(x)} = \frac{\int_{\mathbb{R}} \mu^{-1}\bigl(y\phi_\sigma(y - \mu x) + \phi_\sigma'(y - \mu x)\bigr)\,d\pi(x)}{\int_{\mathbb{R}} \phi_\sigma(y - \mu x)\,d\pi(x)} = \frac{y + \sigma^2(\log p)'(y)}{\mu}, \tag{6.60}$$

so (3.17) holds and g is infinitely differentiable on $\mathbb{R}$, and by similar calculations,

$$g'(y) = \frac{1 + \sigma^2(\log p)''(y)}{\mu} = \frac{\mu}{\sigma^2}\,\frac{\sigma^2\bigl(1 + \sigma^2(\log p)''(y)\bigr)}{\mu^2} = \frac{\mu}{\sigma^2}\,\mathrm{Var}(V_y) \ge 0.$$

We now consider in turn the two conditions on π in the statement of the lemma.

(i) If $V \sim \pi$ has a log-concave density, then the density p of $\mu V + \sigma G$ is also log-concave (Prékopa, 1980), so $(\log p)'' \le 0$ on $\mathbb{R}$. Thus, $0 \le g' \le |\mu|^{-1}$ on $\mathbb{R}$, so g is Lipschitz with constant $|\mu|^{-1}$.

(ii) Suppose first that π is supported on a compact interval $[a, b]$. Then for each $y \in \mathbb{R}$, the distribution π_y has a density with respect to π (by definition), so it is also supported on $[a, b]$. Thus, $\mathrm{Var}(V_y) \le \mathbb{E}\bigl\{\bigl(V_y - (a + b)/2\bigr)^2\bigr\} \le (b - a)^2/4$ for all y, whence g is Lipschitz with constant $|\mu|(b - a)^2/(4\sigma^2)$, and

$$-1 \le \sigma^2(\log p)''(y) \le \frac{\mu^2(b - a)^2}{4\sigma^2} - 1. \tag{6.61}$$

More generally, suppose that π is the distribution of $U_0 + V_0$, where $U_0 \sim N(0, \sigma_0^2)$ with $\sigma_0 \ge 0$, and $V_0 \sim \pi_0$ is independent of U_0 and supported on some compact interval $[a, b]$. Then p is the density of $\mu V + \sigma G \overset{d}{=} \mu U_0 + \sqrt{\sigma^2 + \mu^2\sigma_0^2}\,G$, so it follows from (6.60) and (6.61) that

$$\frac{1}{|\mu|}\biggl(1 - \frac{\sigma^2}{\sigma^2 + \mu^2\sigma_0^2}\biggr) \le |g'| \le \frac{1}{|\mu|}\biggl\{1 + \frac{\sigma^2}{\sigma^2 + \mu^2\sigma_0^2}\biggl(\frac{\mu^2(b - a)^2}{4(\sigma^2 + \mu^2\sigma_0^2)} - 1\biggr)\biggr\}$$

on $\mathbb{R}$. The expression on the right hand side is therefore a Lipschitz constant for g. □

Proof of Lemma 3.7. Let $\psi\colon \mathbb{R}^2 \to [0, \infty)$ be any measurable loss function, and fix $s_1, s_2 \in (0, \infty)$ with $s_1 > s_2$. Taking $G' \sim N(0, s_1^2 - s_2^2)$, $V \sim \pi$ and $G \sim N(0, 1)$ to be jointly independent, we first claim that

$$R_{\pi,\psi}(s_2^{-2}) = \inf_g \mathbb{E}\bigl\{\psi\bigl(g(V + s_2 G), V\bigr)\bigr\} = \inf_{\tilde{g}} \mathbb{E}\bigl\{\psi\bigl(\tilde{g}(V + s_2 G, G'), V\bigr)\bigr\}, \tag{6.62}$$

where the infima are taken over all measurable $g\colon \mathbb{R} \to \mathbb{R}$ and $\tilde{g}\colon \mathbb{R}^2 \to \mathbb{R}$ respectively. Indeed, the first equality holds since $V + s_2 G = s_2(\sqrt{\rho}V + G)$ when $\rho = s_2^{-2}$, and the middle expression is clearly bounded below by the final one, so it remains to prove the reverse inequality. For any fixed $\tilde{g}\colon \mathbb{R}^2 \to \mathbb{R}$, we have

$$\tilde{\Psi}(a) := \mathbb{E}\bigl\{\psi\bigl(\tilde{g}(V + s_2 G, a), V\bigr)\bigr\} \ge \inf_g \mathbb{E}\bigl\{\psi\bigl(g(V + s_2 G), V\bigr)\bigr\} = R_{\pi,\psi}(s_2^{-2})$$

for all $a \in \mathbb{R}$, so it follows from Lemma 7.7 that $\mathbb{E}\bigl\{\psi\bigl(\tilde{g}(V + s_2 G, G'), V\bigr)\bigr\} = \mathbb{E}\bigl(\tilde{\Psi}(G')\bigr) \ge R_{\pi,\psi}(s_2^{-2})$, and hence that (6.62) holds. Since $(V, V + s_2 G + G') \overset{d}{=} (V, V + s_1 G)$, we deduce that

$$\begin{aligned} R_{\pi,\psi}(s_2^{-2}) = \inf_{\tilde{g}} \mathbb{E}\bigl\{\psi\bigl(\tilde{g}(V + s_2 G, G'), V\bigr)\bigr\} &\le \inf_g \mathbb{E}\bigl\{\psi\bigl(g(V + s_2 G + G'), V\bigr)\bigr\} \\ &= \inf_g \mathbb{E}\bigl\{\psi\bigl(g(V + s_1 G), V\bigr)\bigr\} = R_{\pi,\psi}(s_1^{-2}). \end{aligned}$$

In addition, arguing as above for (6.62), we have

$$R_{\pi,\psi}(s_1^{-2}) = \inf_g \mathbb{E}\bigl\{\psi\bigl(g(V + s_1 G), V\bigr)\bigr\} \le \inf_{a \in \mathbb{R}} \mathbb{E}\bigl(\psi(a, V)\bigr) = \inf_g \mathbb{E}\bigl\{\psi\bigl(g(G), V\bigr)\bigr\} = R_{\pi,\psi}(0).$$

Thus, $\rho \mapsto R_{\pi,\psi}(\rho)$ is non-increasing on $[0, \infty)$.

Finally, fix $\rho \in (0,\infty)$ and for each $y \in \mathbb{R}$, let $V_y \sim \pi_y$ be a random variable whose density with respect to π is given by (6.59) with $\mu = \sqrt{\rho}$ and $\sigma = 1$, so that π_y is the conditional (i.e. posterior) distribution of V given $\sqrt{\rho}V + G = y$. It follows from Brown and Purves (1973, Theorem 3) that if the posterior risk function

$$r_y \colon a \mapsto \mathbb{E}\bigl(\psi(a, V_y)\bigr)$$

attains its infimum on $\mathbb{R}$ for Lebesgue almost every $y \in \mathbb{R}$, then there exists a measurable $g^* \equiv g_\rho^* \colon \mathbb{R} \to \mathbb{R}$ with $g^*(y) \in \operatorname{argmin}_{a\in\mathbb{R}} \mathbb{E}\bigl(\psi(a, V_y)\bigr)$ for Lebesgue almost every $y \in \mathbb{R}$, whence $R_{\pi,\psi}(\rho) = \mathbb{E}\bigl\{\psi\bigl(g^*(\sqrt{\rho}V + G), V\bigr)\bigr\}$. This is the case (for every ρ) if $\psi(x,y) = \Psi(x-y)$ for some convex function Ψ with $\Psi(u) \to \infty$ as $|u| \to \infty$, in which case $r_y \colon a \mapsto \mathbb{E}\bigl(\psi(V_y, a)\bigr)$ is convex with $r_y(a) \to \infty$ as $|a| \to \infty$, for each $y \in \mathbb{R}$. □

Corollary 6.19. *Given independent random variables $V \sim \pi$ and $G \sim N(0,1)$, the function $\rho \mapsto \mathrm{mmse}(\rho) := \mathbb{E}\bigl\{\bigl(V - \mathbb{E}(V \,|\, \sqrt{\rho}V + G)\bigr)^2\bigr\}$ is non-increasing on $[0,\infty)$. Moreover, if V satisfies one of the conditions of Lemma 3.8, then $\rho \mapsto \mathrm{mmse}(\rho)$ is also continuous on $(0,\infty)$.*

Proof of Corollary 6.19. Recall that whenever X, Y are random variables with $\mathbb{E}(X^2) < \infty$, it follows from an orthogonal decomposition of the type (3.18) that $\mathbb{E}\bigl\{\bigl(X - \mathbb{E}(X \,|\, Y)\bigr)^2\bigr\} = \min_g \mathbb{E}\bigl\{\bigl(X - g(Y)\bigr)^2\bigr\}$, where the minimum is over all measurable functions $g \colon \mathbb{R} \to \mathbb{R}$. Thus, by Lemma 3.7, $\rho \mapsto \mathrm{mmse}(\rho)$ is non-increasing on $[0,\infty)$.

Now fix $s_1 > s_2 > 0$, and as in the proof of Lemma 3.7, let $G' \sim N(0, s_1^2 - s_2^2)$, $V \sim \pi$ and $G \sim N(0,1)$ be jointly independent, so that $(V, V + s_2 G + G') \overset{d}{=} (V, V + s_1 G)$. Then under the conditions of Lemma 3.8, it follows from (i) and (ii) in its proof that there exists a Lipschitz $g_2^* \colon \mathbb{R} \to \mathbb{R}$ with $g_2^*(V + s_2 G) = \mathbb{E}(V \,|\, V + s_2 G) = \mathbb{E}(V \,|\, s_2^{-1} V + G)$ and Lipschitz constant $L_{s_2} \leq C_\pi (1 \vee s_2^{-2})$, where $C_\pi > 0$ depends only on π. Thus,

$$\begin{aligned}
\mathrm{mmse}(s_2^{-2}) \leq \mathrm{mmse}(s_1^{-2}) &= \min_g \mathbb{E}\bigl\{\bigl(V - g(V + s_1 G)\bigr)^2\bigr\} \\
&\leq \mathbb{E}\bigl\{\bigl(V - g_2^*(V + s_1 G)\bigr)^2\bigr\} \\
&= \mathbb{E}\bigl\{\bigl(V - g_2^*(V + s_2 G + G')\bigr)^2\bigr\} \\
&= \mathbb{E}\bigl\{\bigl(V - g_2^*(V + s_2 G)\bigr)^2\bigr\} + \mathbb{E}\bigl\{\bigl(g_2^*(V + s_2 G) - g_2^*(V + s_2 G + G')\bigr)^2\bigr\} \\
&\leq \mathrm{mmse}(s_2^{-2}) + L_{s_2}\, \mathbb{E}(|G'|^2) \\
&\leq \mathrm{mmse}(s_2^{-2}) + C_\pi\, (s_1^2 s_2^2 \vee s_1^2)(s_2^{-2} - s_1^{-2}).
\end{aligned} \tag{6.63}$$

To justify the equality in the third-last line, note that $g_2^*(V + s_2 G) = \mathbb{E}(V \,|\, V + s_2 G, G')$ by the independence of G' and (V, G), so for any measurable $\tilde{g} \colon \mathbb{R}^2 \to \mathbb{R}$ with $\mathbb{E}\bigl(\tilde{g}(V + s_2 G, G')^2\bigr) < \infty$, we have $\mathbb{E}\bigl\{\bigl(V - g_2^*(V + s_2 G)\bigr)\, \tilde{g}(V + s_2 G, G')\bigr\} = 0$. We deduce from (6.63) that $\rho \mapsto \mathrm{mmse}(\rho)$ is Lipschitz on (ρ', ∞) for every $\rho' > 0$, and hence that it is continuous on $(0,\infty)$. □

Proof of Corollary 3.9. Given any sequence of functions (g_k) for which the corresponding AMP iterations (3.3) satisfy the hypotheses of Theorem 3.1 or 3.5, we prove (3.22) by induction on $k \in \mathbb{N}_0$. For each such k, it follows from (3.11) and (3.19) that as $n \to \infty$, we have

$$\frac{|\langle \hat{v}^k, v\rangle_n|}{\|\hat{v}^k\|_n \|v\|_n} \overset{c}{\to} \frac{\sqrt{\rho_{k+1}}}{\lambda} = \frac{\bigl|\mathbb{E}\bigl(V g_k(\mu_k V + \sigma_k G_k)\bigr)\bigr|}{\sqrt{\mathbb{E}\bigl(g_k(\mu_k V + \sigma_k G_k)^2\bigr)}} \leq \sqrt{\mathbb{E}\bigl(\mathbb{E}(V \,|\, \mu_k V + \sigma_k G_k)^2\bigr)} = \sqrt{1 - \mathrm{mmse}_k(\rho_k)},$$

where we set $\rho_0 \equiv \rho_0^* = (\mu_0/\sigma_0)^2$ and write G_0 for the random variable U from (M1) when $k = 0$. Now $\sqrt{1 - \mathrm{mmse}_0(\rho_0^*)} = \sqrt{\rho_1^*}/\lambda$ by (3.21), so $\rho_1 \leq \rho_1^*$ and (3.22) holds when $k = 0$. For a general $k \in \mathbb{N}$, we have $\rho_k \leq \rho_k^*$ by induction, so since $\rho \mapsto \mathrm{mmse}(\rho) \equiv \mathrm{mmse}_k(\rho)$ is non-increasing by Corollary 6.19, we deduce that $\sqrt{1 - \mathrm{mmse}(\rho_k)} \leq \sqrt{1 - \mathrm{mmse}(\rho_k^*)} = \sqrt{\rho_{k+1}^*}/\lambda$ and hence that $\rho_{k+1} \leq \rho_{k+1}^*$. This completes the inductive step for (3.22).

As for (3.23), we can apply (3.9), the definition of $R_{\pi,\psi}(\rho)$, the fact that $\rho_k \leq \rho_k^*$ and Lemma 3.7 (in that order) to conclude that $n^{-1} \sum_{i=1}^n \psi(\hat{v}_i^k, v_i) \overset{c}{\to} \mathbb{E}\bigl\{\psi\bigl(g_k(\mu_k V + \sigma_k G_k), V\bigr)\bigr\} \geq R_{\psi,\pi}(\rho_k) \geq R_{\psi,\pi}(\rho_k^*)$, as required. □

Proof of Theorem 3.10. Under the conditions of Lemma 3.8, each $g_k^*\colon \mathbb{R} \to \mathbb{R}$ is Lipschitz and satisfies (M2), and by Corollary 6.19, $\rho \mapsto \lambda^2\bigl(1 - \mathrm{mmse}(\rho)\bigr) =: m_\lambda(\rho)$ is non-decreasing on $[0,\infty)$ and continuous on $(0,\infty)$.

(a) We will show that if either (i) or (ii) holds, then

$$\begin{aligned} &\rho_{\mathrm{AMP}}^* \equiv \rho_{\mathrm{AMP}}^*(\lambda) := \inf\{\rho > 0 : \rho = m_\lambda(\rho)\} > 0, && \rho_0^* \in [0, \rho_{\mathrm{AMP}}^*], \qquad \rho_1^* > 0, \\ &m_\lambda(\rho) \geq \rho \text{ for all } \rho \in [0, \rho_{\mathrm{AMP}}^*], && \rho_{k+1}^* = m_\lambda(\rho_k^*) \text{ for all } k \in \mathbb{N}_0. \end{aligned} \tag{6.64}$$

Note that since $m_\lambda(\rho) \leq \lambda^2$ for all $\rho \in [0,\infty)$, we always have $\rho_{\mathrm{AMP}}^*(\lambda) \leq \lambda^2$.

(i) *Non-spectral initialisation*: In this case, (M1) holds with $\mu_0 = 0$, $\sigma_0 = 1$, $\rho_0^* = 0$ and $U = 1$. Since $\mathbb{E}(V) \neq 0$ and $\mathbb{E}(V^2) = 1$ by (M1), $\rho_1^* = \lambda^2\bigl(1 - \mathrm{mmse}_k(0)\bigr) = \lambda^2\bigl(1 - \mathrm{Var}(V)\bigr) = \lambda^2\,\mathbb{E}(V)^2 = m_\lambda(0) > 0$, and therefore $\rho_{k+1}^* = m_\lambda(\rho_k^*)$ for all $k \in \mathbb{N}_0$. In addition, $m_\lambda(\rho) \geq m_\lambda(0) = \rho_1^* > \rho$ for all $\rho \in [0, \rho_1^*)$, so $\rho_{\mathrm{AMP}}^* \geq \rho_1^* > 0$ and $m_\lambda(\rho) > \rho$ for all $\rho \in [0, \rho_{\mathrm{AMP}}^*)$.

(ii) *Spectral initialisation*: By Proposition 3.4, (M1) holds with $U \equiv G_0 \sim N(0,1)$, so $\rho_{k+1}^* = m_\lambda(\rho_k^*)$ for all $k \in \mathbb{N}_0$. Since $\mathrm{mmse}(\rho)$ is the minimum value of $\mathbb{E}\bigl\{\bigl(V - g(\sqrt{\rho}V + G)\bigr)^2\bigr\}$ as g ranges over all measurable functions,

$$m_\lambda(\rho) \geq \lambda^2\Bigl(1 - \inf_{a,b\in\mathbb{R}} \mathbb{E}\bigl\{\bigl(V - a(\sqrt{\rho}V + G) - b\bigr)^2\bigr\}\Bigr) = \lambda^2\left(1 - \frac{1 - \mathbb{E}(V)^2}{1 + \rho\,\{1 - \mathbb{E}(V)^2\}}\right) \geq \frac{\lambda^2\rho}{1+\rho}$$

for all $\rho \in [0,\infty)$, where the equality above can be verified by a routine calculation. Recalling that $\lambda > 1$, we have $m_\lambda(\rho) \geq \lambda^2\rho/(1+\rho) > \rho$ for all $\rho \in [0, \lambda^2 - 1)$, so $\rho_{\mathrm{AMP}}^* \geq \lambda^2 - 1 = \rho_0^* > 0$ and $m_\lambda(\rho) > \rho$ for all $\rho \in (0, \rho_{\mathrm{AMP}}^*)$. Moreover, $\rho_1^* = m_\lambda(\rho_0^*) \geq \rho_0^* > 0$.

To complete the proof of (a), note that if $0 \leq \rho_k^* \leq \rho_{\mathrm{AMP}}^*$ for some $k \in \mathbb{N}_0$, then by (6.64) and the fact that $\rho \mapsto m_\lambda(\rho)$ is non-decreasing, we have $0 \leq \rho_k^* \leq m_\lambda(\rho_k^*) = \rho_{k+1}^* \leq m_\lambda(\rho_{\mathrm{AMP}}^*) = \rho_{\mathrm{AMP}}^*$. Since $\rho_1^* > 0$, it follows by induction that (ρ_k^*) is an increasing sequence that converges to some $\rho^* \in (0, \rho_{\mathrm{AMP}}^*]$. By the continuity of m_λ on $(0,\infty)$, we conclude that $\rho^* = \lim_{k\to\infty} \rho_{k+1}^* = \lim_{k\to\infty} m_\lambda(\rho_k^*) = m_\lambda(\rho^*)$ and hence that $\rho^* = \rho_{\mathrm{AMP}}^*$.

(b) Since $g_{k,\psi}^*$ is Lipschitz by assumption, it follows from Theorem 3.1 that $n^{-1}\sum_{i=1}^n \psi(\hat{v}_i^{k,\psi}, v_i) \xrightarrow{c} R_{\pi,\psi}(\rho_k^*)$; see the proof of Corollary 3.2. Moreover, $\rho_k^* < \rho_{k+1}^*$ by (a) and $R_{\pi,\psi}$ is non-increasing by Lemma 3.7, so $R_{\pi,\psi}(\rho_k^*) \geq R_{\pi,\psi}(\rho_{k+1}^*)$.

(c) Since each g_k^* is Lipschitz and $\rho_k^* \nearrow \rho_{\mathrm{AMP}}^*$ by (a), this is an immediate consequence of (3.10) and (3.11) from Corollary 3.2. □

Proof of Lemma 3.13. This proof proceeds by induction on k. We will first present the argument for the non-spectral initialisation in Theorem 3.10 and then outline the appropriate modifications for the spectral case.

(i) *Non-spectral initialisation*: Defining the state evolution parameters $\mu_k^*, \sigma_k^*, \bar{\Sigma}^{[k]}$ and limiting random variables as in the paragraph containing (3.33), we claim that

$$\mathrm{Cov}\left(\frac{\sigma_k^* G_k}{\mu_k^*}, \frac{\sigma_\ell^* G_\ell}{\mu_\ell^*}\right) = \frac{\bar{\Sigma}_{k,\ell}}{\mu_k^*\,\mu_\ell^*} = \frac{1}{\rho_k^*} \quad \text{and} \quad V \perp\!\!\!\perp (\boldsymbol{\mu}_{\boldsymbol{k-1}}^* V + \bar{\boldsymbol{G}}_{\boldsymbol{k-1}}^*) \mid (\mu_k^* V + \sigma_k^* G_k) \tag{6.65}$$

for all $k \geq \ell \geq 1$, where we use the notation of Section 7.2 to denote conditional independence. For $k = 1$, we have $\mathrm{Var}(\sigma_1^* G_1/\mu_1^*) = (\sigma_1^*/\mu_1^*)^2 = 1/\rho_1^*$. Moreover, $\boldsymbol{\mu}_{\boldsymbol{k-1}}^* V + \bar{\boldsymbol{G}}_{\boldsymbol{k-1}}^* = \mu_0 V + \sigma_0 U = cU$ by condition (i) of Theorem 3.10, and $U \perp\!\!\!\perp (V,\, \mu_1^* V + \sigma_1^* G_1)$ by definition in (3.31). Therefore, $V \perp\!\!\!\perp cU \mid (\mu_1^* V + \sigma_1^* G_1)$, which verifies (6.65) when $k = 1$.

For a general $k \geq 2$ and $1 \leq \ell \leq k$, it follows from (3.31) and (3.33) that

$$\begin{aligned}\bar{\Sigma}_{k,\ell} &= \mathbb{E}\big(\boldsymbol{g}_{\boldsymbol{\ell-1}}^*(\boldsymbol{\mu}_{\boldsymbol{\ell-1}}^* V + \bar{\boldsymbol{G}}_{\boldsymbol{\ell-1}}^*) \cdot \boldsymbol{g}_{\boldsymbol{k-1}}^*(\boldsymbol{\mu}_{\boldsymbol{k-1}}^* V + \bar{\boldsymbol{G}}_{\boldsymbol{k-1}}^*)\big) \\ &= \mathbb{E}\big(\mathbb{E}(V \,|\, \boldsymbol{\mu}_{\boldsymbol{\ell-1}}^* V + \bar{\boldsymbol{G}}_{\boldsymbol{\ell-1}}^*) \cdot \mathbb{E}(V \,|\, \boldsymbol{\mu}_{\boldsymbol{k-1}}^* V + \bar{\boldsymbol{G}}_{\boldsymbol{k-1}}^*)\big) \\ &= \mathbb{E}\big(V\mathbb{E}(V \,|\, \boldsymbol{\mu}_{\boldsymbol{\ell-1}}^* V + \bar{\boldsymbol{G}}_{\boldsymbol{\ell-1}}^*)\big) = \mu_\ell^*/\lambda \\ &= \mathbb{E}\big(\mathbb{E}(V \,|\, \boldsymbol{\mu}_{\boldsymbol{\ell-1}}^* V + \bar{\boldsymbol{G}}_{\boldsymbol{\ell-1}}^*)^2\big) = (\sigma_\ell^*)^2,\end{aligned} \tag{6.66}$$

where we have used the tower property of conditional expectation to obtain the last two equalities above. Thus, $\rho_k^* = (\mu_k^*/\sigma_k^*)^2 = \lambda\mu_k^*$ and

$$\mathrm{Cov}\bigg(\frac{\sigma_k^* G_k}{\mu_k^*}, \frac{\sigma_\ell^* G_\ell}{\mu_\ell^*}\bigg) = \frac{\bar{\Sigma}_{k,\ell}}{\mu_k^* \, \mu_\ell^*} = \frac{1}{\lambda\mu_k^*} = \frac{1}{\rho_k^*}. \tag{6.67}$$

Now for $1 \leq \ell \leq k$, let $Z_\ell := \sigma_\ell^* G_\ell/\mu_\ell^*$ and $\zeta_\ell := Z_\ell - Z_k$, so that $\mathrm{Cov}(\zeta_\ell, Z_k) = 0$ by (6.67). Since $(\zeta_1, \ldots, \zeta_{k-1}, Z_k)$ is a Gaussian random vector that is independent of (U, V), it follows that $(\zeta_1, \ldots, \zeta_{k-1})$, Z_k, U and V are mutually independent, whence $V \perp\!\!\!\perp (cU, \zeta_1, \ldots, \zeta_{k-1}, V{+}Z_k) \mid (V{+}Z_k)$ by Lemma 7.8. Writing $\mu_\ell^* V + \sigma_\ell^* G_\ell = \mu_\ell^*(V + Z_\ell) = \mu_\ell^*\big((V + Z_k) + \zeta_\ell\big)$ for $1 \leq \ell \leq k$, we deduce that

$$V \perp\!\!\!\perp (cU, \mu_1^* V + \sigma_1^* G_1, \ldots, \mu_{k-1}^* V + \sigma_{k-1}^* G_{k-1}) \mid (\mu_k^* V + \sigma_k^* G_k),$$

which completes the inductive step for (6.65). An immediate consequence of this conditional independence is (3.34), which implies that $\boldsymbol{g}_{\boldsymbol{k}}^*$ depends only its last argument and $b_{kj}^* = \langle \partial_j \boldsymbol{g}_{\boldsymbol{k}}^*(v^0, \ldots, v^k)\rangle_n = 0$ for $1 \leq j \leq k-1$. Thus, with the denoising functions $\boldsymbol{g}_{\boldsymbol{k}}^*$ given by (3.34), we see by induction that the state evolution recursion (3.31) reduces to that in (3.20), and hence that (3.30) coincides with the Bayes-AMP iteration $(v^{k,\mathrm{B}} \equiv v^{k,\mathrm{B}}(n) : k, n \in \mathbb{N})$ in Section 3.3.

(ii) *Spectral initialisation*: In this case, the initial state evolution parameters are $\mu_0^* = \sqrt{1 - \lambda^2}$ and $\sigma_0^* = 1/\lambda$. We will show by induction that (6.65) holds for all $k \geq \ell \geq 0$, with the convention $\boldsymbol{\mu}_{\boldsymbol{-1}}^* V + \bar{\boldsymbol{G}}_{\boldsymbol{-1}}^* = 0$. The base case $k = 0$ is trivial since $\mathrm{Var}(\sigma_0^* G_0/\mu_0^*) = (\sigma_0^*/\mu_0^*)^2 = 1/\rho_0^* = 1/(\lambda^2 - 1)$ and the second part of (6.65) holds vacuously. For a general $k \in \mathbb{N}$, (6.66) once again holds for all $1 \leq \ell \leq k$, and for $\ell = 0$, the appropriate generalisation of the first line of (3.15) yields

$$\begin{aligned}\bar{\Sigma}_{k,0} &= \lambda^{-1}\,\mathbb{E}\big((\mu_0^* V + \sigma_0^* G_0) \cdot \boldsymbol{g}_{\boldsymbol{k-1}}^*(\boldsymbol{\mu}_{\boldsymbol{k-1}}^* V + \bar{\boldsymbol{G}}_{\boldsymbol{k-1}})\big) \\ &= \lambda^{-1}\,\mathbb{E}\big((\mu_0^* V + \sigma_0^* G_0) \cdot \mathbb{E}(V \,|\, \boldsymbol{\mu}_{\boldsymbol{k-1}}^* V + \bar{\boldsymbol{G}}_{\boldsymbol{k-1}}^*)\big) \\ &= \lambda^{-1}\,\mathbb{E}\big(V(\mu_0^* V + \sigma_0^* G_0)\big) = \mu_0^*/\lambda,\end{aligned}$$

where we have used the tower property of conditional expectation in the second equality. It then follows that (6.67) holds for all $0 \leq \ell \leq k$. Defining Z_ℓ and ζ_ℓ as above for $0 \leq \ell \leq k$, we deduce that $V \perp\!\!\!\perp (\zeta_0, \zeta_1, \ldots, \zeta_{k-1}, V + Z_k) \mid (V + Z_k)$, and the rest of the argument is essentially the same as in (i). □

6.9 Proofs for Section 4

The proof of Lemma 4.1 makes use of the following multivariate version of Stein's lemma.

Lemma 6.20. *Let $g\colon \mathbb{R}^d \to \mathbb{R}$ be such that for $j = 1, \ldots, d$, the function $x_j \mapsto g(x_1, \ldots, x_d)$ is absolutely continuous for Lebesgue almost every $(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d) \in \mathbb{R}^{d-1}$, with weak derivative $\partial_j g\colon \mathbb{R}^d \to \mathbb{R}$ satisfying $\mathbb{E}\big(|\partial_j g(X)|\big) < \infty$. Let $\nabla g(x) := (\partial_1(x), \ldots, \partial_d(x))$ for $x \in \mathbb{R}^d$. If $X \sim N_d(0, \Sigma)$ with Σ positive definite, then*

$$\mathbb{E}\big(Xg(X)\big) = \Sigma\,\mathbb{E}\big(\nabla g(X)\big).$$

Proof. The result for $\Sigma = I_d$ is stated as Tsybakov (2009, Lemma 3.6). For a general non-negative definite $\Sigma \in \mathbb{R}^{d\times d}$, let $\tilde{g}(z) = g(\Sigma^{1/2}z)$ for $z \in \mathbb{R}^d$. Then $\nabla\tilde{g}(z) = \Sigma^{1/2}\,\nabla g(\Sigma^{1/2}z)$ for all z (Fourdrinier et al., 2018, Theorem 2.1; Fan, 2022, Proposition E.5), so by taking $Z \sim N_d(0, I_d)$, we conclude that

$$\mathbb{E}\big(Xg(X)\big) = \Sigma^{1/2}\,\mathbb{E}\big(Z\tilde{g}(Z)\big) = \Sigma^{1/2}\,\mathbb{E}\big(\nabla\tilde{g}(Z)\big) = \Sigma\,\mathbb{E}\big(\nabla g(X)\big). \qquad \square$$

Proof of Lemma 4.1. The first assertion follows from the following standard fact: let (X_1, X_2) be a Gaussian random vector with $\mathbb{E}(X_1) = \mathbb{E}(X_2) = 0$ and $\Sigma_{ij} := \mathrm{Cov}(X_i, X_j)$ for $i, j \in \{1, 2\}$. If $\Sigma_{11} = \mathrm{Var}(X_1)$ is invertible, then AX_1 and $X_2 - AX_1$ are uncorrelated and hence independent when $A := \Sigma_{21}\Sigma_{11}^{-1}$. Thus, $(X_1, X_2) \overset{d}{=} (X_1, AX_1 + G)$, where $AX_1 = \mathbb{E}(X_2 \,|\, X_1)$ and $G \sim N(0, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$ is independent of X_1. For (4.8), we deduce from Lemma 6.20 that

$$\mathbb{E}\big(Z\tilde{g}_k(Z, Z_k, v)\big) = \Sigma_{11}\,\mathbb{E}\big(\partial_1\tilde{g}_k(Z, Z_k, v)\big) + \Sigma_{12}\,\mathbb{E}\big(\partial_2\tilde{g}_k(Z, Z_k, v)\big)$$

for all $v \in \mathbb{R}$, where $\Sigma \equiv \Sigma_k$ is as in (4.7). Thus, since $(Z, Z_k) \overset{d}{=} (Z, \mu_{Z,k}Z + \sigma_{Z,k}\tilde{G}_k)$ is independent of $\bar{\varepsilon}$ in (4.6), and $\mu_{Z,k} = \Sigma_{21}/\Sigma_{11}$ by the first part of the lemma,

$$\begin{aligned}\mathbb{E}\big(Zg_k(Z_k, Y)\big) = \mathbb{E}\big(Z\tilde{g}_k(Z, Z_k, \bar{\varepsilon})\big) &= \Sigma_{11}\,\mathbb{E}\big(\partial_1\tilde{g}_k(Z, Z_k, \bar{\varepsilon})\big) + \Sigma_{12}\,\mathbb{E}\big(\partial_2\tilde{g}_k(Z, Z_k, \bar{\varepsilon})\big) \\ &= \frac{\mathbb{E}(\bar{\beta}^2)}{\delta}\big\{\mu_{k+1} + \mu_{Z,k}\,\mathbb{E}\big(g_k'(Z_k, Y)\big)\big\},\end{aligned} \tag{6.68}$$

which yields the first equality. Next, by the tower property of expectation, the final expression in (4.9) can be written as

$$\mathbb{E}\bigg(\frac{\mathbb{E}(Z \,|\, Z_k, Y) - \mathbb{E}(Z \,|\, Z_k)}{\mathrm{Var}(Z \,|\, Z_k)}\, g_k(Z_k, Y)\bigg) = \mathbb{E}\bigg(\frac{Z - \mathbb{E}(Z \,|\, Z_k)}{\mathrm{Var}(Z \,|\, Z_k)}\, \tilde{g}_k(Z, Z_k, \bar{\varepsilon})\bigg).$$

Since Z is conditionally Gaussian given Z_k and $\bar{\varepsilon}$ is independent of (Z, Z_k), a further (conditional) application of Stein's lemma yields

$$\mathbb{E}\bigg(\frac{Z - \mathbb{E}(Z \,|\, Z_k)}{\mathrm{Var}(Z \,|\, Z_k)}\, \tilde{g}_k(Z, Z_k, \bar{\varepsilon}) \;\bigg|\; Z_k\bigg) = \frac{\mathbb{E}\big\{\big(Z - \mathbb{E}(Z \,|\, Z_k)\big)\, \tilde{g}_k(Z, Z_k, \bar{\varepsilon}) \,\big|\, Z_k\big\}}{\mathrm{Var}(Z \,|\, Z_k)} = \mathbb{E}\big(\partial_1\tilde{g}_k(Z, Z_k, \bar{\varepsilon}) \,\big|\, Z_k\big),$$

so by taking expectations, we obtain the second identity for μ_{k+1}. □

Proof of Proposition 4.4. Consider the right hand side of (4.31) and write $\bar{J}(\tilde{\beta}) := \sum_{j=1}^p J(\tilde{\beta}_j)$ for $\tilde{\beta} \in \mathbb{R}^p$. Using the expression for the Lagrangian (4.30), and ignoring terms that do not depend on $\tilde{\beta}$, we obtain

$$\begin{aligned}\operatorname*{argmin}_{\tilde{\beta}\in\mathbb{R}^p}\Big\{L(\tilde{\beta}, \hat{\theta}^k, \hat{s}^k) - \frac{\bar{c}_k}{2}\|\tilde{\beta} - \hat{\beta}^k\|^2\Big\} &= \operatorname*{argmin}_{\tilde{\beta}\in\mathbb{R}^p}\Big\{\bar{J}(\tilde{\beta}) - \tilde{\beta}^\top X^\top \hat{s}^k - \frac{\bar{c}_k}{2}\|\tilde{\beta} - \hat{\beta}^k\|^2\Big\} \\ &= \operatorname*{argmin}_{\tilde{\beta}\in\mathbb{R}^p}\bigg\{\bar{J}(\tilde{\beta}) - \frac{\bar{c}_k}{2}\bigg\|\tilde{\beta} + \frac{X^\top\hat{s}^k - \bar{c}_k\hat{\beta}^k}{\bar{c}_k}\bigg\|^2\bigg\} \\ &= \operatorname*{argmin}_{\tilde{\beta}\in\mathbb{R}^p}\bigg\{\bar{J}(\tilde{\beta}) - \frac{\bar{c}_k}{2}\bigg\|\tilde{\beta} + \frac{\beta^{k+1}}{\bar{c}_k}\bigg\|^2\bigg\} = \hat{\beta}^{k+1},\end{aligned} \tag{6.69}$$

where the third and final equalities follow from the definitions of β^{k+1} and $\hat{\beta}^{k+1} = f_{k+1}(\beta^{k+1})$ respectively in (4.29), with f_{k+1} as in (4.24). Similarly, in view of the definition of $\bar{g}_k$ in (4.23), we can obtain (4.32) by completing the square. For (4.33), we can apply (4.29) to see that

$$\hat{s}^{k+1} = \frac{\hat{\theta}^{k+1} - \theta^{k+1}}{\bar{b}_{k+1}} = \frac{\hat{\theta}^{k+1} - (X\hat{\beta}^{k+1} - \bar{b}_{k+1}\hat{s}^k)}{\bar{b}_{k+1}} = \hat{s}^k + \frac{(\hat{\theta}^{k+1} - X\hat{\beta}^{k+1})}{\bar{b}_{k+1}}. \tag{6.70}$$

For the final assertion of Proposition 4.4, if $(\beta^*, \theta^*, \hat{\beta}^*, \hat{\theta}^*, \hat{s}^*)$ is a fixed point of the algorithm (4.29), then $\hat{\theta}^* = X\hat{\beta}^*$ by (4.33), and (for example by considering subgradients) it follows from (4.31) and (4.32) respectively that

$$\begin{aligned}\hat{\beta}^* &= \operatorname*{argmin}_{\tilde{\beta}\in\mathbb{R}^p} L(\tilde{\beta}, \hat{\theta}^*, \hat{s}^*) = \operatorname*{argmin}_{\tilde{\beta}\in\mathbb{R}^p}\big\{\bar{J}(\tilde{\beta}) - (X\tilde{\beta})^\top\hat{s}^*\big\}, \\ \hat{\theta}^* &= \operatorname*{argmin}_{\tilde{\theta}\in\mathbb{R}^n} L(\hat{\beta}^*, \tilde{\theta}, \hat{s}^*) = \operatorname*{argmin}_{\tilde{\theta}\in\mathbb{R}^n}\big\{\bar{\ell}(\tilde{\theta}, y) + \tilde{\theta}^\top\hat{s}^*\big\},\end{aligned}$$

where $\bar{\ell}(\tilde{\theta}, y) := \sum_{i=1}^n \ell(\tilde{\theta}_i, y_i)$. Thus, for all $(\tilde{\beta}, \tilde{\theta}) \in \mathbb{R}^p \times \mathbb{R}^n$ with $\tilde{\theta} = X\tilde{\beta}$, we have

$$\begin{aligned}\bar{J}(\tilde{\beta}) + \bar{\ell}(\tilde{\theta}, y) &= \big(\bar{J}(\tilde{\beta}) - (X\tilde{\beta})^\top\hat{s}^*\big) + \big(\bar{\ell}(\tilde{\theta}, y) + \tilde{\theta}^\top\hat{s}^*\big) \\ &\geq \big(\bar{J}(\hat{\beta}^*) - (X\hat{\beta}^*)^\top\hat{s}^*\big) + \big(\bar{\ell}(\hat{\theta}^*, y) + (\hat{\theta}^*)^\top\hat{s}^*\big) = \bar{J}(\hat{\beta}^*) + \bar{\ell}(\hat{\theta}^*, y),\end{aligned}$$

so $(\hat{\beta}^*, \hat{\theta}^*)$ is a solution to the optimisation problem (4.22), as required. □

7 Supplementary mathematical background

7.1 Basic properties of complete convergence

Proof of Proposition 1.2. For (a), suppose that $\sum_n \mathbb{P}(\|X_n\|_E > \varepsilon) < \infty$ for all $\varepsilon > 0$. Then for any sequence (Y_n) of E-valued random elements with $Y_n \overset{d}{=} X_n$ for all n, the first Borel–Cantelli lemma implies that $\mathbb{P}(\|Y_n\|_E > \varepsilon$ infinitely often$) = 0$ for all $\varepsilon > 0$ and hence that $Y_n \to 0$ almost surely. This shows that $X_n \xrightarrow{c} 0$. Conversely, suppose that $\sum_n \mathbb{P}(\|X_n\|_E > \varepsilon) = \infty$ for some $\varepsilon > 0$. Then for a sequence (Y_n) of independent E-valued random elements with $Y_n \overset{d}{=} X_n$ for all n, the second Borel–Cantelli lemma implies that $\mathbb{P}(\|Y_n\|_E > \varepsilon$ infinitely often$) = 1$ and hence that $Y_n \nrightarrow 0$ almost surely. Thus, $X_n \overset{c}{\nrightarrow} 0$.

The argument for (b) is similar. If $\sum_n \mathbb{P}(\|X_n\|_E > C) < \infty$ for all $C > 0$, then $X_n = O_c(1)$ by the first Borel–Cantelli lemma. Conversely, suppose that $\sum_n \mathbb{P}(\|X_n\|_E > C) = \infty$ for all $C > 0$. Then for a sequence (Y_n) of independent E-valued random elements with $Y_n \overset{d}{=} X_n$ for all n, the second Borel–Cantelli lemma implies that $\mathbb{P}(\|Y_n\|_E > C$ infinitely often$) = 1$ for all $C > 0$ and hence that $\limsup_{n\to\infty} \|Y_n\|_E = \infty$ almost surely. Thus, (X_n) is not $O_c(1)$. □

Remark 7.1. For a random sequence (X_n) taking values in a Euclidean space $(E, \|\cdot\|_E)$, it can be seen from Definition 1.1 and Proposition 1.2 that complete convergence (to a degenerate limit) is a property of the marginal distributions of the random elements $X_1, X_2, \ldots$ and not of their joint dependence structure (i.e. the specific coupling between them), so $X_1, X_2, \ldots$ need not be defined on the same probability space. Thus, just as for weak convergence or convergence in probability to a degenerate limit (but not almost sure convergence), there is a meaningful notion of complete convergence for sequences (μ_n) of Borel probability measures on E: defining $\bar{B}(x, \varepsilon) := \{x' \in E : \|x' - x\|_E \le \varepsilon\}$ for $\varepsilon > 0$, we write $\mu_n \xrightarrow{c} \delta_x$ if $\sum_n \mu_n\bigl(\bar{B}(x, \varepsilon)^c\bigr) < \infty$ for all $\varepsilon > 0$.

Example 1. Let (X_n) be any sequence of random variables for which there exist $c_1, c_2, \beta > 0$ such that $\mathbb{P}(|X_n| > t) \le c_1 \exp(-c_2 t^\beta)$ for all $t > 0$ and $n \in \mathbb{N}$. Let (a_n) be a deterministic sequence of real numbers. If $a_n = o(1)$, then clearly $a_n X_n = o_p(1)$, and if $a_n = O(1)$, then $a_n X_n = O_p(1)$. Moreover:

(a) If $|a_n|^\beta \log n \to 0$, then for every $t > 0$, there exists $N \in \mathbb{N}$ such that $c_2(t/|a_n|)^\beta \ge 2 \log n$ for all $n > N$, so $\sum_n \mathbb{P}(|a_n X_n| > t) \le N + \sum_{n>N} c_1 e^{-2\log n} < \infty$. Thus, $a_n X_n = o_c(1)$ by Proposition 1.2(a).

(b) If $\limsup_{n\to\infty} |a_n|^\beta \log n < \infty$, then there exists $t > 0$ and $N \in \mathbb{N}$ such that $c_2(t/|a_n|)^\beta \ge 2 \log n$ for all $n > N$, so $\sum_n \mathbb{P}(|a_n X_n| > t) < \infty$ as in (i). Thus, $a_n X_n = O_c(1)$ by Proposition 1.2(b).

Suppose in addition that there exist $c_1', c_2', \beta' > 0$ such that $\mathbb{P}(|X_n| > t) \ge c_1' \exp(-c_2' t^{\beta'})$ for all $t > 0$ and $n \in \mathbb{N}$.

(c) If $\liminf_{n\to\infty} |a_n|^{\beta'} \log n > 0$, then there exist $t > 0$ and $N \in \mathbb{N}$ such that $c_2'(t/|a_n|)^{\beta'} \le \log n$ for all $n > N$, so $\sum_n \mathbb{P}(|a_n X_n| > t) \ge \sum_{n\ge N} c_1 e^{-\log n} = \infty$. Thus, $(a_n X_n)$ is not $o_c(1)$ in view of Proposition 1.2(a).

(d) If $|a_n|^{\beta'} \log n \to \infty$, then for every $t > 0$, there exists $N \in \mathbb{N}$ such that $c_2'(t/|a_n|)^{\beta'} \le \log n$ for all $n > N$, so $\sum_n \mathbb{P}(|a_n X_n| > t) = \infty$ as in (iii). Thus, $(a_n X_n)$ is not $O_c(1)$ in view of Proposition 1.2(b).

For instance, suppose that $X_n = X \sim N(0, 1)$ for all n. Then $X_n \to X$ almost surely and $X_n = O_p(1)$ but (X_n) is not $O_c(1)$, and $X_n / \log^{1/2} n \to 0$ almost surely but $(X_n / \log^{1/2} n)$ is not $o_c(1)$.

Using Proposition 1.2, it is straightforward to verify that the continuous mapping theorem and Slutsky's lemma remain valid when stated in terms of complete convergence.

Lemma 7.2. *Let $(X_n), (Y_n)$ be sequences of random elements taking values in Euclidean spaces E, E' respectively such that $X_n \overset{c}{\to} x$ and $Y_n \overset{c}{\to} y$ for some deterministic limits $x \in E$ and $y \in E'$. Then $(X_n, Y_n) \overset{c}{\to} (x, y)$ in $E \times E'$ and $g(X_n) \overset{c}{\to} g(x)$ in E' for any function $g\colon E \to E'$ that is continuous at x. If in addition x lies in some open set $U \subseteq E$, then $\mathbb{1}_{\{X_n \notin U\}} \overset{c}{\to} 0$.*

Consequently, $X_n + Y_n \overset{c}{\to} x + y$ when $E = E'$. Moreover, $X_n Y_n \overset{c}{\to} xy$ when $E = \mathbb{R}$ and E' is any Euclidean space (in the case of scalar multiplication), or when $E = \mathbb{R}^{k \times \ell}$ and $E' = \mathbb{R}^{\ell}$ for some $k, \ell \in \mathbb{N}$ (in the case of matrix multiplication). If in addition $k = \ell$ and $x^{-1} \in E$ is well-defined, then $\mathbb{1}_{\{X_n \text{ is not invertible}\}} \overset{c}{\to} 0$ and $X_n^{+} Y_n \overset{c}{\to} x^{-1}y$.

Proof. For the first part of the lemma, we apply Proposition 1.2(a). Since $\|(\tilde{x}, \tilde{y})\|_{E \times E'}^2 = \|\tilde{x}\|_E^2 + \|\tilde{y}\|_{E'}^2$ for any $(\tilde{x}, \tilde{y}) \in E \times E'$, we have

$$\sum_n \mathbb{P}\big(\|(X_n, Y_n) - (x, y)\|_{E \times E'} > \varepsilon\big) \leq \sum_n \big\{\mathbb{P}(\|X_n - x\|_E > \varepsilon/\sqrt{2}) + \mathbb{P}(\|Y_n - y\|_{E'} > \varepsilon/\sqrt{2})\big\} < \infty$$

for all $\varepsilon > 0$, so $(X_n, Y_n) \overset{c}{\to} (x, y)$ by Proposition 1.2(a). If $g\colon E \to E'$ is continuous at $x \in E$, then for each $\varepsilon > 0$, there exists $\delta > 0$ such that $\|g(\tilde{x}) - g(x)\|_{E'} < \varepsilon$ whenever $\|\tilde{x} - x\|_E < \delta$, so

$$\sum_n \mathbb{P}(\|g(X_n) - g(x)\|_{E'} > \varepsilon) \leq \sum_n \mathbb{P}(\|X_n - x\|_E > \delta) < \infty.$$

This holds for all $\varepsilon > 0$, so $g(X_n) \overset{c}{\to} g(x)$ by Proposition 1.2(a). When x lies in some open set $U \subseteq E$, there exists $\varepsilon > 0$ such that $\mathbb{P}(X_n \notin U) \leq \mathbb{P}(\|X_n - X\|_E > \varepsilon)$, so $\mathbb{1}_{\{X_n \notin U\}} \overset{c}{\to} 0$, again by Proposition 1.2(a).

Having established the first part of the lemma, we can now apply the facts above to deduce the remaining assertions. Indeed, when $E = E'$, the function $g\colon (\tilde{x}, \tilde{y}) \mapsto \tilde{x} + \tilde{y}$ is continuous on $E \times E'$ and we know that $(X_n, Y_n) \overset{c}{\to} (x, y)$, so it follows that $X_n + Y_n \overset{c}{\to} x + y$. When $E = \mathbb{R}$ or when $E = \mathbb{R}^{k \times \ell}$ and $E' = \mathbb{R}^{\ell}$ for some $k, \ell \in \mathbb{N}$, the scalar and matrix multiplication maps (respectively) are continuous on $E \times E'$. Therefore, it follows similarly that $X_n Y_n \overset{c}{\to} xy$.

If in addition $k = \ell$, then $\tilde{x}^{+} = \tilde{x}^{-1}$ for all invertible $\tilde{x} \in E = \mathbb{R}^{k \times k}$, so the map $\tilde{x} \mapsto \tilde{x}^{+}$ is continuous on the set U of all invertible $\tilde{x} \in E$, which is open. A further application of the continuous mapping result above shows that if x is invertible, then $\mathbb{1}_{\{X_n \text{ is not invertible}\}} \overset{c}{\to} 0$ and $X_n^{+} Y_n \overset{c}{\to} x^{-1}y$, as required. □

Remark 7.3. By a similar application of Proposition 1.2, it can be shown that the stochastic o_c and O_c symbols obey the 'arithmetic rules' of standard O notation. Written in compact form, some examples of these are as follows (for sequences defined on spaces with compatible dimensions):

$$\begin{aligned} O_c(1) + O_c(1) &= O_c(1), & o_c(1) + o_c(1) &= o_c(1), & O_c(1) + o_c(1) &= O_c(1), \\ O_c(1)\, O_c(1) &= O_c(1), & o_c(1)\, o_c(1) &= o_c(1), & O_c(1)\, o_c(1) &= O_c(1), \end{aligned} \tag{7.1}$$

where the assertions in the second line apply to scalar multiplication or matrix multiplication as appropriate. (The proofs are straightforward and are therefore omitted.) To give another example of a basic fact that follows directly from Definition 1.1 or Proposition 1.2, let E, E' be Euclidean spaces and suppose that $g\colon E \to E'$ is bounded on every bounded subset of E. Then for any sequence (X_n) of E-valued random elements such that $X_n = O_c(1)$, we also have $g(X_n) = O_c(1)$.

7.2 Regular conditional distributions and conditional independence

First, we recall the notion of conditional expectation: if $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $\mathcal{G} \subseteq \mathcal{F}$ is a sub-σ-algebra, we write $\mathbb{P}|_{\mathcal{G}}$ for the restricted probability measure on $(\Omega, \mathcal{G})$ given by $\mathbb{P}|_{\mathcal{G}}(B) := \mathbb{P}(B)$ for $B \in \mathcal{G}$. If $Y\colon (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ is a random variable with $\mathbb{E}(|Y|) < \infty$, then there exists a $\mathcal{G}$-measurable random variable $Z = \mathbb{E}(Y \,|\, \mathcal{G})$ with the property that $\mathbb{E}(Z\mathbb{1}_E) = \mathbb{E}(Y\mathbb{1}_E)$ for all $E \in \mathcal{G}$

(Dudley, 2002, Chapter 10.1). We call Z the *conditional expectation of* Y *given* $\mathcal{G}$, noting that it is unique up to $\mathbb{P}|_{\mathcal{G}}$-almost sure equivalence. For $F \in \mathcal{F}$, we also write $\mathbb{P}(F \,|\, \mathcal{G}) := \mathbb{E}(\mathbb{1}_F \,|\, \mathcal{G})$.

If X is a measurable function from $(\Omega, \mathcal{F}, \mathbb{P})$ to a measurable space $(\mathcal{X}, \mathcal{A})$, we say that $P_{X|\mathcal{G}} \colon \Omega \times \mathcal{A} \to [0,1]$ is a *(regular) conditional distribution for* X *given* $\mathcal{G}$ if

(i) for every $\omega \in \Omega$, the set function $P_{X|\mathcal{G}}(\omega, \cdot)$ is a probability measure on $\mathcal{A}$;

(ii) for each $A \in \mathcal{A}$, the map $P_{X|\mathcal{G}}(\cdot, A)$ is $\mathcal{G}$-measurable, and $P_{X|\mathcal{G}}(\omega, A) = \mathbb{P}(X^{-1}(A) \,|\, \mathcal{G})(\omega)$ for $\mathbb{P}|_{\mathcal{G}}$-almost every $\omega \in \Omega$, so that $\mathbb{P}(X^{-1}(A) \cap E) = \int_E P_{X|\mathcal{G}}(\omega, A)\, d\mathbb{P}(\omega)$ for all $E \in \mathcal{G}$.

We say that $(\mathcal{X}, \mathcal{A})$ is a *Borel space* if there exist a Borel subset $S \subseteq [0,1]$ (equipped with the restriction $\mathcal{B}_S$ of the Borel σ-algebra on $[0,1]$ to S) and a bijection $f \colon (\mathcal{X}, \mathcal{A}) \to (S, \mathcal{B}_S)$ such that both f and f^{-1} are measurable. Examples of Borel spaces $(\mathcal{X}, \mathcal{A})$ include *Polish spaces* (i.e. separable, completely metrisable topological spaces) $\mathcal{X}$ equipped with their Borel σ-algebras $\mathcal{A}$ (Kallenberg, 1997, Theorem A1.6).

Whenever $(\mathcal{X}, \mathcal{A})$ is a Borel space, there exists a conditional distribution $P_{X|\mathcal{G}}$, and moreover, if P' is another such conditional distribution, then $P'(\omega, \cdot) = P_{X|\mathcal{G}}(\omega, \cdot)$ for $\mathbb{P}|_{\mathcal{G}}$-almost every $\omega \in \Omega$; see Kallenberg (1997, Theorem 5.3) and Dudley (2002, Theorem 10.2.2). For brevity, we will write 'X has conditional distribution $P \equiv P_\omega$ given $\mathcal{G}$ on an event $\Omega_0 \in \mathcal{G}$' to mean that there exists a conditional distribution $P_{X|\mathcal{G}}$, and we can take $P_{X|\mathcal{G}}(\omega, \cdot) = P_\omega(\cdot)$ for $\mathbb{P}|_{\mathcal{G}}$-almost every $\omega \in \Omega_0$. When we omit the phrase 'on an event $\Omega_0 \in \mathcal{G}$', we mean that the statement holds for $\mathbb{P}|_{\mathcal{G}}$-almost every $\omega \in \Omega$.

For measurable $X, X' \colon (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{X}, \mathcal{A})$, we say that X, X' are *identically distributed given* $\mathcal{G}$, and write $X \overset{d}{=}|_{\mathcal{G}} X'$, if there exist conditional distributions $P_{X|\mathcal{G}}$ and $P'_{X'|\mathcal{G}}$ for X, X' respectively, and $P_\omega(\cdot) \equiv P_{X|\mathcal{G}}(\omega, \cdot) = P'_{X'|\mathcal{G}}(\omega, \cdot) \equiv P'_\omega(\cdot)$ for $\mathbb{P}|_{\mathcal{G}}$-almost every $\omega \in \Omega$.

Remark 7.4. For example, X has distribution Q on $(\mathcal{X}, \mathcal{A})$ and is independent of $\mathcal{G}$ if and only if X has conditional distribution $P_\omega = Q$ for all $\omega \in \Omega$.

Remark 7.5. Let $X \colon (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{X}, \mathcal{A})$ be as above and consider the important special case where $\mathcal{G} = \sigma(Y)$ for some measurable map Y from $(\Omega, \mathcal{F}, \mathbb{P})$ to a measurable space $(\mathcal{Y}, \mathcal{B})$. Denote by P the joint distribution of $(X, Y) \colon (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B})$ and by P^Y the (marginal) distribution of Y on $(\mathcal{Y}, \mathcal{B})$. We note here that a random variable $Z \colon (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ is $\sigma(Y)$-measurable if and only if $Z = g \circ Y$ for some measurable function $g \colon (\mathcal{Y}, \mathcal{B}) \to \mathbb{R}$ (i.e. '$Z(\omega)$ depends on ω only through $Y(\omega)$'); see for example Dudley (2002, Theorem 4.2.8). Using this fact and the defining property (ii) above, it can be verified (as in Dudley, 2002, Theorem 10.2.1) that there exists a regular conditional distribution $P_{X|\sigma(Y)} \colon \Omega \times \mathcal{A} \to [0,1]$ if and only if there is a family of probability distributions $(Q_y)_{y \in \mathcal{Y}}$ on $(\mathcal{X}, \mathcal{A})$ such that the following hold for every $A \in \mathcal{A}$:

(I) $y \mapsto Q_y(A)$ is a measurable function from $(\mathcal{Y}, \mathcal{B})$ to $\mathbb{R}$;

(II) $P(A \times B) = \mathbb{P}\big(X^{-1}(A) \cap Y^{-1}(B)\big) = \int_B Q_y(A)\, dP^Y(y)$ for all $B \in \mathcal{B}$.

In this case, for $\mathbb{P}|_{\sigma(Y)}$-almost every $\omega \in \Omega$, we have $P_{X|\sigma(Y)}(\omega, A) = Q_{Y(\omega)}(A)$ for all $A \in \mathcal{A}$. Note that $(Q_y)_{y \in \mathcal{Y}}$ is only unique up to P^Y-almost sure equivalence, in the sense that if $(Q'_y)_{y \in \mathcal{Y}}$ satisfies (I) and $Q_y = Q'_y$ for P^Y-almost every $y \in \mathcal{Y}$, then $(Q'_y)_{y \in \mathcal{Y}}$ also satisfies (II). In view of (I), the map $(y, A) \mapsto Q_y(A)$ is said to be a *probability kernel*. An interpretation of (II) is that it makes precise the notion of *disintegrating* the joint distribution P of (X, Y) into the marginal distribution P^Y of Y and the distributions $(Q_y)_{y \in \mathcal{Y}}$, where (for P^Y-almost every $y \in \mathcal{Y}$) we can view Q_y as the "conditional distribution of X given $Y = y$". Indeed, by analogy with the construction of the usual product measure and Fubini's theorem (e.g. Dudley, 2002, Chapter 4.4), it can be shown that if $\phi \colon (\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B}) \to \mathbb{R}$ is P-integrable (i.e. ϕ is measurable and $\mathbb{E}\big(|\phi(X, Y)|\big) < \infty$), then

(III) $x \mapsto \phi(x, y)$ is $\mathcal{A}$-measurable for all $y \in \mathcal{Y}$ and Q_y-integrable for P^Y-almost every $y \in \mathcal{Y}$;

(IV) $y \mapsto \int_{\mathcal{X}} \phi(x,y)\,dQ_y(x)$ is $\mathcal{B}$-measurable and P^Y-integrable;

(V) $\int_{\mathcal{X}\times\mathcal{Y}} \phi(x,y)\,dP(x,y) = \int_{\mathcal{Y}} \bigl(\int_{\mathcal{X}} \phi(x,y)\,dQ_y(x)\bigr)\,dP^Y(y)$.

This generalisation of Fubini's theorem is sometimes known as the *disintegration theorem*, and is derived from (II) using a monotone class argument; see Dudley (2002, Theorem 10.2.1(II)) and Kallenberg (1997, Theorem 5.4).

Lemma 7.6. *Let $(\mathcal{X},\mathcal{A}),(\mathcal{Y},\mathcal{B})$ be measurable spaces and let $(\mathcal{Z},\mathcal{C})$ be a Borel space. Let $\phi\colon(\mathcal{X}\times\mathcal{Y},\mathcal{A}\otimes\mathcal{B})\to(\mathcal{Z},\mathcal{C})$ be a measurable function and let $\mathcal{G}\subseteq\mathcal{F}$ be a σ-algebra.*

(a) If $E\in\mathcal{G}$ and $X_1,X_2\colon(\Omega,\mathcal{F},\mathbb{P})\to(\mathcal{X},\mathcal{A})$ are measurable functions with conditional distributions $P\equiv P_\omega$ and $Q\equiv Q_\omega$ respectively given $\mathcal{G}$, then the measurable function $X\colon(\Omega,\mathcal{F},\mathbb{P})\to(\mathcal{X},\mathcal{A})$ satisfying $X=X_1$ on E and $X=X_2$ on E^c has conditional distribution $R(\cdot)\equiv R_\omega(\cdot):=P_\omega(\cdot)\mathbb{1}_{\{\omega\in E\}}+Q_\omega(\cdot)\mathbb{1}_{\{\omega\in E^c\}}$ given $\mathcal{G}$.

(b) For $D\in\mathcal{A}\otimes\mathcal{B}$ and $y\in\mathcal{Y}$, let $D^y:=\{x\in\mathcal{X}:(x,y)\in D\}=\iota_y^{-1}(D)$, where $\iota_y\colon\mathcal{X}\to\mathcal{X}\times\mathcal{Y}$ denotes the map $x\mapsto(x,y)$. Fix $\Omega_0\in\mathcal{G}$. Suppose that $X\colon(\Omega,\mathcal{F},\mathbb{P})\to(\mathcal{X},\mathcal{A})$ has conditional distribution $P\equiv P_\omega$ given $\mathcal{G}$ on Ω_0, and that $Y\colon(\Omega,\mathcal{F},\mathbb{P})\to(\mathcal{Y},\mathcal{B})$ is $\mathcal{G}$-measurable. If $Z\colon(\Omega,\mathcal{F},\mathbb{P})\to(\mathcal{Z},\mathcal{C})$ is a measurable map that agrees with $\phi(X,Y)$ on Ω_0, then Z has conditional distribution $\tilde{P}\equiv\tilde{P}_\omega=P_\omega\circ(\phi\circ\iota_{Y(\omega)})^{-1}$ given $\mathcal{G}$ on Ω_0, so that $\tilde{P}_\omega(C)=P_\omega\bigl(\phi^{-1}(C)^{Y(\omega)}\bigr)$ for all $C\in\mathcal{C}$ and $\omega\in\Omega_0$.

(c) Suppose that $X,X'\colon(\Omega,\mathcal{F},\mathbb{P})\to(\mathcal{X},\mathcal{A})$ are measurable functions satisfying $X\overset{d}{=}|_{\mathcal{G}}\,X'$, and that $Y\colon(\Omega,\mathcal{F},\mathbb{P})\to(\mathcal{Y},\mathcal{B})$ is $\mathcal{G}$-measurable. Then $\phi(X,Y)\overset{d}{=}|_{\mathcal{G}}\,\phi(X',Y)$.

The result in (b) has an intuitive interpretation. Suppose for simplicity that $\Omega_0=\Omega$, and fix $\omega\in\Omega$. Let $\mu:=P_\omega$ be taken from the conditional distribution of X given $\mathcal{G}$, and assume that Y is $\mathcal{G}$-measurable. To obtain the corresponding $\tilde{P}_\omega$ from the conditional distribution of $\phi(X,Y)$ given $\mathcal{G}$, Lemma 7.6(b) tells us that we can take $\tilde{P}_\omega$ to be the distribution of $\phi(U,y)$, where $U\sim\mu$ and $y:=Y(\omega)$. In essence, the reason for this is that since Y is $\mathcal{G}$-measurable, we can think of Y as being 'fixed' once we have conditioned on $\mathcal{G}$.

Proof. (a) The fact that $R_\omega(\cdot)$ is a probability measure on $\mathcal{A}$ for $\mathbb{P}|_{\mathcal{G}}$-almost every $\omega\in\Omega$ follows immediately from the corresponding facts for $P_\omega(\cdot)$ and $Q_\omega(\cdot)$. For each $A\in\mathcal{A}$, the map $\omega\mapsto R_\omega(A)$ is a composition of $\mathcal{G}$-measurable functions (since $E\in\mathcal{G}$ by assumption), so is $\mathcal{G}$-measurable.

For $A\in\mathcal{A}$, let $\chi_A\colon\mathcal{X}\to\{0,1\}$ denote the indicator function of A. Then

$$\chi_A\circ X=(\chi_A\circ X_1)\mathbb{1}_E+(\chi_A\circ X_2)\mathbb{1}_{E^c}.$$

Since $\mathbb{1}_E$ and $\mathbb{1}_{E^c}$ are $\mathcal{G}$-measurable, it follows that

$$\begin{aligned}\mathbb{P}\bigl(X^{-1}(A)\bigm|\mathcal{G}\bigr)(\omega)&=\mathbb{E}\bigl((\chi_A\circ X_1)\mathbb{1}_E\bigm|\mathcal{G}\bigr)(\omega)+\mathbb{E}\bigl((\chi_A\circ X_2)\mathbb{1}_{E^c}\bigm|\mathcal{G}\bigr)(\omega)\\&=\mathbb{E}\bigl(\chi_A\circ X_1\bigm|\mathcal{G}\bigr)(\omega)\mathbb{1}_E(\omega)+\mathbb{E}\bigl((\chi_A\circ X_2)\bigm|\mathcal{G}\bigr)(\omega)\mathbb{1}_{E^c}(\omega)\\&=P_\omega(A)\mathbb{1}_E(\omega)+Q_\omega(A)\mathbb{1}_{E^c}(\omega)=R_\omega(A)\end{aligned}$$

for $\mathbb{P}|_{\mathcal{G}}$-almost every $\omega\in\Omega$, as required.

(b) This can be deduced from Kallenberg (1997, Theorem 5.4) and part (a) above, but we give a direct proof here for completeness. Note that for $\mathbb{P}|_{\mathcal{G}}$-almost every $\omega\in\Omega_0$, the set function $\tilde{P}_\omega$ is the push-forward (image measure) of P_ω induced by the measurable map $x\mapsto\phi\circ\iota_{Y(\omega)}(x)$ from $(\mathcal{X},\mathcal{A})$ to $(\mathcal{Z},\mathcal{C})$; thus, $\tilde{P}_\omega$ is indeed a probability measure for $\mathbb{P}|_{\mathcal{G}}$-almost every $\omega\in\Omega_0$.

Now let $\mathcal{D}$ denote the collection of all $D\in\mathcal{A}\otimes\mathcal{B}$ for which $\omega\mapsto P_\omega\bigl(D^{Y(\omega)}\bigr)$ is $\mathcal{G}$-measurable and $P_\omega\bigl(D^{Y(\omega)}\bigr)=\mathbb{P}\bigl((X,Y)^{-1}(D)\bigm|\mathcal{G}\bigr)(\omega)$ for $\mathbb{P}|_{\mathcal{G}}$-almost every $\omega\in\Omega_0$. If $D=A\times B$ for some $A\in\mathcal{A}$ and

$B \in \mathcal{B}$, then $D^{Y(\omega)} = A$ if $Y(\omega) \in B$, and $D^{Y(\omega)} = \emptyset$ if $Y(\omega) \notin B$. Thus,

$$\begin{aligned} P_\omega\bigl(D^{Y(\omega)}\bigr) &= P_\omega(A)\mathbb{1}_{\{Y(\omega)\in B\}} = \mathbb{P}\bigl(X^{-1}(A) \bigm| \mathcal{G}\bigr)(\omega)\mathbb{1}_{\{Y(\omega)\in B\}} \\ &= \mathbb{E}\bigl(\chi_A \circ X \bigm| \mathcal{G}\bigr)(\omega) \cdot (\chi_B \circ Y)(\omega) = \mathbb{E}\bigl((\chi_A \circ X) \cdot (\chi_B \circ Y) \bigm| \mathcal{G}\bigr)(\omega) \\ &= \mathbb{P}\bigl((X,Y)^{-1}(D) \bigm| \mathcal{G}\bigr)(\omega) \end{aligned}$$

for $\mathbb{P}|_{\mathcal{G}}$-almost every $\omega \in \Omega_0$, where we have used the fact that $\chi_B \circ Y$ is $\mathcal{G}$-measurable in the penultimate equality. Thus $\mathcal{D} \supseteq \{A \times B : A \in \mathcal{A},\, B \in \mathcal{B}\}$, which is a π-system that generates $\mathcal{D}$. Now suppose that $D_1, D_2 \in \mathcal{D}$ with $D_1 \subseteq D_2$. Then

$$\begin{aligned} P_\omega\bigl((D_2 \setminus D_1)^{Y(\omega)}\bigr) &= P_\omega\bigl(D_2^{Y(\omega)} \setminus D_1^{Y(\omega)}\bigr) = P_\omega\bigl(D_2^{Y(\omega)}\bigr) - P_\omega\bigl(D_1^{Y(\omega)}\bigr) \\ &= \mathbb{P}\bigl((X,Y)^{-1}(D_2) \bigm| \mathcal{G}\bigr)(\omega) - \mathbb{P}\bigl((X,Y)^{-1}(D_1) \bigm| \mathcal{G}\bigr)(\omega) \\ &= \mathbb{P}\bigl((X,Y)^{-1}(D_2 \setminus D_1) \bigm| \mathcal{G}\bigr)(\omega) \end{aligned}$$

for $\mathbb{P}|_{\mathcal{G}}$-almost every $\omega \in \Omega_0$, so $D_2 \setminus D_1 \in \mathcal{D}$. Finally, let (D_n) be an increasing sequence of sets in $\mathcal{D}$, and let $D := \bigcup_{n=1}^\infty D_n$. Then $D^{Y(\omega)} = \bigcup_{n=1}^\infty D_n^{Y(\omega)}$ and $(X,Y)^{-1}(D) = \bigcup_{n=1}^\infty (X,Y)^{-1}(D_n)$, so that

$$P_\omega\bigl(D^{Y(\omega)}\bigr) = \lim_{n\to\infty} P_\omega\bigl(D_n^{Y(\omega)}\bigr) = \lim_{n\to\infty} \mathbb{P}\bigl((X,Y)^{-1}(D_n) \bigm| \mathcal{G}\bigr)(\omega) = \mathbb{P}\bigl((X,Y)^{-1}(D) \bigm| \mathcal{G}\bigr)(\omega)$$

for $\mathbb{P}|_{\mathcal{G}}$-almost every $\omega \in \Omega_0$, where we have used the conditional monotone convergence theorem in the final equality (Dudley, 2002, Theorem 10.1.7). Thus, $D \in \mathcal{D}$, and it follows from Dynkin's lemma that $\mathcal{D} = \mathcal{A} \otimes \mathcal{B}$.

Finally, if $C \in \mathcal{C}$, then $D := \phi^{-1}(C) \in \mathcal{A} \otimes \mathcal{B}$, so that for $\mathbb{P}|_{\mathcal{G}}$-almost every $\omega \in \Omega_0$,

$$\begin{aligned} P_\omega\bigl(\phi^{-1}(C)^{Y(\omega)}\bigr) &= P_\omega\bigl(D^{Y(\omega)}\bigr) = \mathbb{P}\bigl((X,Y)^{-1}(D) \bigm| \mathcal{G}\bigr)(\omega) \\ &= \mathbb{P}\bigl((X,Y)^{-1}(D) \bigm| \mathcal{G}\bigr)(\omega) \cdot \mathbb{1}_{\Omega_0}(\omega) = \mathbb{P}\bigl((X,Y)^{-1}(D) \cap \Omega_0 \bigm| \mathcal{G}\bigr)(\omega) \\ &= \mathbb{P}\bigl(Z^{-1}(C) \cap \Omega_0 \bigm| \mathcal{G}\bigr)(\omega) = \mathbb{P}\bigl(Z^{-1}(C) \bigm| \mathcal{G}\bigr)(\omega), \end{aligned}$$

as required, since $\Omega_0 \in \mathcal{G}$.

(c) This follows directly from (b) on setting $\Omega_0 = \Omega$. □

The following useful result is a special case of Kallenberg (1997, Theorem 5.4) and can be derived using the definition of conditional expectation (Dudley, 2002, Problem 10.1.9), or alternatively using regular conditional distributions and standard measure-theoretic devices (similarly to the proofs of Lemma 7.6(b) above and Dudley (2002, Theorem 10.2.5)).

Lemma 7.7. *Let X, Y be measurable functions from $(\Omega, \mathcal{F}, \mathbb{P})$ to measurable spaces $(\mathcal{X}, \mathcal{A}), (\mathcal{Y}, \mathcal{B})$ respectively, and let $\phi\colon (\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B}) \to \mathbb{R}$ be a measurable function satisfying $\mathbb{E}\bigl(|\phi(X,Y)|\bigr) < \infty$. Let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra, and suppose that Y is $\mathcal{G}$-measurable. If X has distribution Q on $(\mathcal{X}, \mathcal{A})$ and is independent of $\mathcal{G}$, then $\mathbb{E}\bigl(\phi(X,Y) \,\big|\, \mathcal{G}\bigr)(\omega) = \int_{\mathcal{X}} \phi\bigl(x, Y(\omega)\bigr)\, dQ(x)$ for $\mathbb{P}|_{\mathcal{G}}$-almost every $\omega \in \Omega$.*

Next, for σ-algebras $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3 \subseteq \mathcal{F}$, we say that $\mathcal{G}_1$ and $\mathcal{G}_2$ are *conditionally independent given* $\mathcal{G}_3$, and write $\mathcal{G}_1 \perp\!\!\!\perp \mathcal{G}_2 \mid \mathcal{G}_3$, if $\mathbb{P}(A_1 \cap A_2 \,|\, \mathcal{G}_3) = \mathbb{P}(A_1 \,|\, \mathcal{G}_3)\mathbb{P}(A_2 \,|\, \mathcal{G}_3)$ almost surely for all $A_1 \in \mathcal{G}_1$ and $A_2 \in \mathcal{G}_2$, or equivalently if $\mathbb{P}\bigl(A_1 \,|\, \sigma(\mathcal{G}_2, \mathcal{G}_3)\bigr) = \mathbb{P}(A_1 \,|\, \mathcal{G}_3)$ almost surely for all $A_1 \in \mathcal{G}_1$ (Kallenberg, 1997, Proposition 5.6). If this holds with $\mathcal{G}_1 = \sigma(X)$ for some random variable X, we also say that X and $\mathcal{G}_2$ are conditionally independent given $\mathcal{G}_3$, and write $X \perp\!\!\!\perp \mathcal{G}_2 \mid \mathcal{G}_3$ (and similarly for $\mathcal{G}_2$ and $\mathcal{G}_3$). The following basic facts follow straightforwardly from the definition of conditional independence.

Lemma 7.8 (Kallenberg, 1997, Corollary 5.7(i))**.** *We have $\mathcal{G}_1 \perp\!\!\!\perp \mathcal{G}_2 \mid \mathcal{G}_3$ if and only if $\sigma(\mathcal{G}_1, \mathcal{G}_3) \perp\!\!\!\perp \mathcal{G}_2 \mid \mathcal{G}_3$.*

Lemma 7.9. *Let $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$ be measurable spaces and let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra.*

(a) For $i = 1, 2$, let $X_i\colon (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{X}, \mathcal{A})$ and $Y_i\colon (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{Y}, \mathcal{B})$ be measurable functions such that $X_i \perp\!\!\!\perp Y_i \mid \mathcal{G}$. For $E \in \mathcal{G}$, let $X\colon (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{X}, \mathcal{A})$ be the measurable function satisfying $X = X_1$ on E and $X = X_2$ on E^c, and define Y similarly. Then $X \perp\!\!\!\perp Y \mid \mathcal{G}$.

(b) *Suppose that the measurable maps* $X\colon(\Omega,\mathcal{F},\mathbb{P})\to(\mathcal{X},\mathcal{A})$ *and* $Y\colon(\Omega,\mathcal{F},\mathbb{P})\to(\mathcal{Y},\mathcal{B})$ *have conditional distributions* $P\equiv P_\omega$ *and* $Q\equiv Q_\omega$ *respectively given* $\mathcal{G}$*. Then* $X\perp\!\!\!\perp Y\mid\mathcal{G}$ *if and only if* $(X,Y)\colon(\Omega,\mathcal{F},\mathbb{P})\to(\mathcal{X}\times\mathcal{Y},\mathcal{A}\otimes\mathcal{B})$ *has conditional distribution* $R\equiv R_\omega:=P_\omega\otimes Q_\omega$ *given* $\mathcal{G}$*.*

Proof. (a) For fixed $A\in\mathcal{A}$ and $B\in\mathcal{B}$, write $\chi_A\colon(\mathcal{X},\mathcal{A})\to\{0,1\}$ and $\chi_B\colon(\mathcal{Y},\mathcal{B})\to\{0,1\}$ for the respective indicator functions, and for $i=1,2$, note that

$$\begin{aligned}\mathbb{E}\big((\chi_A\circ X_i)\cdot(\chi_B\circ Y_i)\bigm|\mathcal{G}\big)=\mathbb{P}\big(X_i^{-1}(A)\cap Y_i^{-1}(B)\bigm|\mathcal{G}\big)&=\mathbb{P}\big(X_i^{-1}(A)\bigm|\mathcal{G}\big)\mathbb{P}\big(Y_i^{-1}(B)\bigm|\mathcal{G}\big)\\&=\mathbb{E}(\chi_A\circ X_i\,|\,\mathcal{G})\,\mathbb{E}(\chi_B\circ Y_i\,|\,\mathcal{G})\end{aligned}$$

almost surely, since $X_i\perp\!\!\!\perp Y_i\mid\mathcal{G}$. As in the proof of Lemma 7.6(a), we have $\chi_A\circ X=(\chi_A\circ X_1)\mathbb{1}_E+(\chi_{A^c}\circ X_2)\mathbb{1}_{E^c}$ and $\chi_B\circ Y=(\chi_B\circ Y_1)\mathbb{1}_E+(\chi_{B^c}\circ Y_2)\mathbb{1}_{E^c}$, so it follows that

$$\begin{aligned}\mathbb{P}\big(X^{-1}(A)\cap Y^{-1}(B)\mid\mathcal{G}\big)&=\mathbb{E}\big((\chi_A\circ X)\cdot(\chi_B\circ Y)\bigm|\mathcal{G}\big)\\&=\mathbb{E}\big((\chi_A\circ X_1)\cdot(\chi_B\circ Y_1)\mathbb{1}_E+(\chi_A\circ X_2)\cdot(\chi_B\circ Y_2)\mathbb{1}_{E^c}\bigm|\mathcal{G}\big)\\&=\mathbb{E}\big((\chi_A\circ X_1)\cdot(\chi_B\circ Y_1)\bigm|\mathcal{G}\big)\mathbb{1}_E+\mathbb{E}\big((\chi_A\circ X_2)\cdot(\chi_B\circ Y_2)\bigm|\mathcal{G}\big)\mathbb{1}_{E^c}\\&=\mathbb{E}(\chi_A\circ X_1\,|\,\mathcal{G})\,\mathbb{E}(\chi_B\circ Y_1\,|\,\mathcal{G})\mathbb{1}_E+\mathbb{E}(\chi_A\circ X_2\,|\,\mathcal{G})\,\mathbb{E}(\chi_B\circ Y_2\,|\,\mathcal{G})\mathbb{1}_{E^c}\\&=\mathbb{E}\big((\chi_A\circ X_1)\mathbb{1}_E+(\chi_A\circ X_2)\mathbb{1}_{E^c}\bigm|\mathcal{G}\big)\,\mathbb{E}\big((\chi_B\circ Y_1)\mathbb{1}_E+(\chi_B\circ Y_2)\mathbb{1}_{E^c}\,|\,\mathcal{G}\big)\\&=\mathbb{E}(\chi_A\circ X\,|\,\mathcal{G})\,\mathbb{E}(\chi_B\circ Y\,|\,\mathcal{G})=\mathbb{P}\big(X^{-1}(A)\bigm|\mathcal{G}\big)\,\mathbb{P}\big(Y^{-1}(B)\bigm|\mathcal{G}\big),\end{aligned}$$

where we have used the fact that $E\in\mathcal{G}$ to obtain the third-last equality. Since this holds for all $A\in\mathcal{A}$ and $B\in\mathcal{B}$, the result follows.

(b) For $A\in\mathcal{A}$ and $B\in\mathcal{B}$, note that

$$\mathbb{P}\big(X^{-1}(A)\cap Y^{-1}(B)\bigm|\mathcal{G}\big)(\omega)=\mathbb{P}\big((X,Y)^{-1}(A\times B)\bigm|\mathcal{G}\big)(\omega)\tag{7.2}$$

$$\mathbb{P}\big(X^{-1}(A)\bigm|\mathcal{G}\big)(\omega)\cdot\mathbb{P}\big(Y^{-1}(B)\bigm|\mathcal{G}\big)(\omega)=P_\omega(A)\,Q_\omega(B)=R_\omega(A\times B)\tag{7.3}$$

for $\mathbb{P}|_{\mathcal{G}}$-almost every $\omega\in\Omega$. Thus, if (X,Y) has conditional distribution $R\equiv R_\omega=P_\omega\otimes Q_\omega$ given $\mathcal{G}$, then for any $A\in\mathcal{A}$ and $B\in\mathcal{B}$, the right hand sides of (7.2) and (7.3) agree for $\mathbb{P}|_{\mathcal{G}}$-almost every $\omega\in\Omega$, so the same is true of the left hand sides. This shows that $X\perp\!\!\!\perp Y\mid\mathcal{G}$.

Conversely, suppose that $X\perp\!\!\!\perp Y\mid\mathcal{G}$ and let $\mathcal{D}$ be the collection of all $D\in\mathcal{A}\otimes\mathcal{B}$ such that $R_\omega(D)=(P_\omega\otimes Q_\omega)(D)=\mathbb{P}\big((X,Y)^{-1}(D)\bigm|\mathcal{G}\big)(\omega)$ for $\mathbb{P}|_{\mathcal{G}}$-almost every $\omega\in\Omega$. Then for any $A\in\mathcal{A}$ and $B\in\mathcal{B}$, the left hand sides of (7.2) and (7.3) agree for $\mathbb{P}|_{\mathcal{G}}$-almost every $\omega\in\Omega$, so $\mathcal{D}$ contains a π-system $\{A\times B:A\in\mathcal{A},\,B\in\mathcal{B}\}$ that generates $\mathcal{A}\otimes\mathcal{B}$. Similarly to the proof of Lemma 7.6(b), it can be verified that $\mathcal{D}$ is a d-system, so it follows from Dynkin's lemma that $\mathcal{D}=\mathcal{A}\otimes\mathcal{B}$, and hence that (X,Y) has conditional distribution $R\equiv R_\omega=P_\omega\otimes Q_\omega$, as required. □

7.3 Auxiliary probabilistic results

The following general result is used in the proofs of some important complete convergence statements in Sections 6.4 and 7.4, specifically Proposition 6.16(g) and Corollary 7.21(b).

Lemma 7.10. *Let* $(X_n),(Y_n)$ *be sequences of measurable functions defined on* $(\Omega,\mathcal{F},\mathbb{P})$ *such that* X_n,Y_n *take values in Polish spaces* E_n,E_n' *respectively for each* $n\in\mathbb{N}$*. Suppose that there exist Borel measurable functions* $g_n\colon E_n\to E_n'$ *such that* $Y_n\overset{d}{=}g_n(X_n)$ *for each* n*. Then there exists a sequence of measurable functions* $\tilde{X}_n\colon\Omega\to E_n$ *such that* $\tilde{X}_n\overset{d}{=}X_n$ *for all* n *and* $\big(g_1(\tilde{X}_1),g_2(\tilde{X}_2),\dots\big)=(Y_1,Y_2,\dots)$ *almost surely (viewed as random sequences taking values in* $\prod_{n=1}^\infty E_n'$*, equipped with its cylindrical (i.e. Borel)* σ*-algebra).*

This is an extension to random sequences of the following result for pairs of random elements: given random elements X_1,X_2 taking values in E_1,E_2 respectively, let $(Y_1,Y_2)\sim\pi$ be any coupling of $g_1(X_1),g_2(X_2)$. Then there exists a coupling $(X_1',X_2')\sim\pi'$ of X_1,X_2 such that $\big(g_1(X_1'),g_2(X_2')\big)\overset{d}{=}$

(Y_1, Y_2), i.e. $\pi = \pi' \circ (g_1, g_2)^{-1}$. This can be proved by applying the gluing lemma from optimal transport (Villani, 2003, Lemma 7.6) or a simpler version of the general argument below.

Given an arbitrary coupling $(Y_1, Y_2, \dots)$ of the random elements $g_1(X_1), g_2(X_2), \dots$, the first (and most important) step in the proof below is to 'lift' this to produce a suitable coupling $(X_1', X_2', \dots)$ of the random elements $X_1, X_2, \dots$, in such a way that $\bigl(g_1(X_1'), g_2(X_2'), \dots\bigr) \overset{d}{=} (Y_1, Y_2, \dots)$ as random sequences. Intuitively, the key construction can be interpreted as the output of the following two-stage procedure:

(A) Denoting by π the (given) distribution of $(Y_1, Y_2, \dots)$ on $\prod_{n=1}^\infty E_n'$, we first draw $(Y_1', Y_2', \dots) \sim \pi$;

(B) Having obtained $(Y_1', Y_2', \dots) = (y_1, y_2, \dots)$ from Step A, we then generate $X_1', X_2', \dots$ by sampling independently from $Q_{y_1}^1, Q_{y_2}^2, \dots$, where $Q_{y_n}^n$ denotes the "conditional distribution of X_n given $g_n(X_n) = y_n$".

Step B ensures that $X_1', X_2', \dots$ are conditionally independent given $(Y_1', Y_2', \dots)$. To make rigorous sense of this informal description and to validate the construction, we use the language of *disintegration of measures*, as outlined in Remark 7.5. There are similarities here with the proof of the gluing lemma (Villani, 2003, Lemma 7.6). To verify that the random sequences $\bigl(g_n(X_n')\bigr)$ and (Y_n) have the same distribution on $\prod_{n=1}^\infty E_n'$, it suffices to show that they have the same finite-dimensional distributions, i.e. that $\bigl(g_1(X_1'), \dots, g_n(X_n')\bigr) \overset{d}{=} (Y_1, \dots, Y_n)$ for all n. Finally, to upgrade all the distributional equalities above to almost-sure equalities, we appeal to a general result from abstract probability theory (Kallenberg, 1997, Corollary 5.11), which is also proved using disintegration techniques.

Remark 7.11. To guarantee the existence of a random sequence $(\tilde{X}_1, \tilde{X}_2, \dots)$ with a given distribution on $\prod_{n=1}^\infty E_n'$, we require the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to be rich enough to support a sequence of independent $U[0,1]$ random variables. This can be assumed without loss of generality, since otherwise we can work with the product space $(\Omega \times [0,1], \mathcal{F} \otimes \mathcal{B}_{[0,1]}, \mathbb{P} \otimes \mu_{[0,1]})$, where $\mathcal{B}_{[0,1]}$ and $\mu_{[0,1]}$ denote the Borel σ-algebra and Lebesgue measure on $[0,1]$ respectively.

Proof of Lemma 7.10. For each n, denote by $\mathcal{B}_n, \mathcal{B}_n'$ the Borel σ-algebras of E_n, E_n' respectively. It follows from Dudley (2002, Theorem 2.5.7) and Kallenberg (1997, Lemma 1.2) that $\prod_{j=1}^n E_j'$ and $\prod_{j=1}^n E_j$ are Polish spaces with Borel σ-algebras $\bigotimes_{j=1}^n \mathcal{B}_j'$ and $\bigotimes_{j=1}^n \mathcal{B}_j$ respectively. Denote by μ_n, π_n the distributions of Y_n and $(Y_1, \dots, Y_n)$ on $(E_n', \mathcal{B}_n')$ and $\bigl(\prod_{j=1}^n E_j', \bigotimes_{j=1}^n \mathcal{B}_j'\bigr)$ respectively. Since E_n is a Polish space, we know from Section 7.2 that there exists a regular conditional distribution for X_n given $\sigma\bigl(g_n(X_n)\bigr)$. Equivalently, there is a family of probability distributions $(Q_y^n)_{y \in E_n'}$ on E_n satisfying conditions (I) and (II) in Remark 7.5, where we take $X := X_n$, $Y := g_n(X_n) \overset{d}{=} Y_n$ and $P^Y := \mu_n$. It follows from Remark 7.5(II) that $\mathbb{P}(X_n \in A) = \int_{E_n'} Q_y^n(A)\, d\mu_n(y)$ for all $A \in \mathcal{B}_n$, and moreover that

$$\int_{B'} \mathbb{1}_B(y)\, d\mu_n(y) = \mathbb{P}\bigl(g_n(X_n) \in B \cap B'\bigr) = \mathbb{P}\bigl(X_n \in g_n^{-1}(B),\, g_n(X_n) \in B'\bigr) = \int_{B'} Q_y^n\bigl(g_n^{-1}(B)\bigr)\, d\mu_n(y) \tag{7.4}$$

for $B, B' \in \mathcal{B}_n'$. Thus, for all $B \in \mathcal{B}_n'$, we have $Q_y^n\bigl(g_n^{-1}(B)\bigr) = \mathbb{1}_B(y)$ for μ_n-almost every $y \in E_n'$.

For each $n \in \mathbb{N}$, we now define a new measure π_n' on $\bigl(\prod_{j=1}^n E_j, \bigotimes_{j=1}^n \mathcal{B}_j\bigr)$ by

$$\pi_n'(A) := \int_{\prod_{j=1}^n E_j'} \biggl(\int_{E_1} \cdots \int_{E_n} \mathbb{1}_A(x_1, \dots, x_n)\, dQ_{y_n}^n(x_n) \cdots dQ_{y_1}^1(x_1)\biggr)\, d\pi_n(y_1, \dots, y_n) \tag{7.5}$$

for $A \in \bigotimes_{j=1}^n \mathcal{B}_j$. That this a well-defined probability measure follows from Remark 7.5(III, IV) and the monotone convergence theorem. For each n, we claim that

(i) $\pi_{n+1}'(A \times E_{n+1}) = \pi_n'(A)$ for every $A \in \bigotimes_{j=1}^n \mathcal{B}_j$;

(ii) $\mathbb{P}(X_n \in A_n) = \pi_n'\bigl(\prod_{j=1}^{n-1} E_j \times A_n\bigr)$ for every $A_n \in \mathcal{B}_n$;

(iii) $\pi_n = \pi'_n \circ (g_1,\dots,g_n)^{-1}$, where $(g_1,\dots,g_n)\colon \bigl(\prod_{j=1}^n E_j, \bigotimes_{j=1}^n \mathcal{B}_j\bigr) \to \bigl(\prod_{j=1}^n E'_j, \bigotimes_{j=1}^n \mathcal{B}'_j\bigr)$ denotes the measurable map $(x_1,\dots,x_n) \mapsto \bigl(g_1(x_1),\dots,g_n(x_n)\bigr)$.

Property (i) is immediate from (7.5) and the fact that $\pi_{n+1}(B \times E'_n) = \pi_n(B)$ for all $B \in \bigotimes_{j=1}^n \mathcal{B}'_j$. To verify (ii), observe that

$$\begin{aligned}\pi'_n\bigl(\textstyle\prod_{j=1}^{n-1} E_j \times A_n\bigr) &= \int_{\prod_{j=1}^n E'_j} \int_{E_n} \mathbb{1}_{A_n}(x_n)\, dQ^n_{y_n}(x_n)\, d\pi_n(y_1,\dots,y_n)\\ &= \int_{E'_n} Q^n_{y_n}(A_n)\, d\mu_n(y_n) = \mathbb{P}(X_n \in A_n),\end{aligned}$$

where the final equality is obtained from Remark 7.5(II) as above. As for (iii), fix $B_j \in \mathcal{B}'_j$ for $1 \le j \le n$ and note that by (7.4) and (7.5), we have

$$\begin{aligned}\pi'_n\bigl(\textstyle\prod_{j=1}^n g_j^{-1}(B_j)\bigr) &= \int_{\prod_{j=1}^n E'_j} Q^1_{y_1}\bigl(g_1^{-1}(B_1)\bigr)\cdots Q^n_{y_n}\bigl(g_n^{-1}(B_n)\bigr)\, d\pi_n(y_1,\dots,y_n)\\ &= \int_{\prod_{j=1}^n E'_j} \mathbb{1}_{B_1}(y_1)\cdots\mathbb{1}_{B_n}(y_n)\, d\pi_n(y_1,\dots,y_n) = \pi_n\bigl(\textstyle\prod_{j=1}^n B_j\bigr).\end{aligned}$$

This means that π_n and $\pi'_n \circ (g_1,\dots,g_n)^{-1}$ agree on $\bigl\{\prod_{j=1}^n B_j : B_j \in \mathcal{B}'_j \text{ for all } 1 \le j \le n\bigr\}$, a π-system that generates $\bigotimes_{j=1}^n \mathcal{B}_j$, so (iii) holds.

Since the distributions $\pi'_1, \pi'_2, \dots$ on the Polish spaces $E_1, E_1 \times E_2, \dots$ satisfy the consistency condition (i), we deduce from the Daniell–Kolmogorov extension theorem (Kallenberg, 1997, Theorem 5.14) and Remark 7.11 that exists a sequence $(X'_n)_{n\in\mathbb{N}}$ of random elements $X'_n\colon \Omega \to E_n$ such that $(X'_1,\dots,X'_n) \sim \pi'_n$ on $\prod_{j=1}^n E_j$ for each n. Then by (ii) and (iii) above, we have $X'_n \overset{d}{=} X_n$ and $\bigl(g_1(X'_1),\dots,g_n(X'_n)\bigr) \sim \pi'_n \circ (g_1,\dots,g_n)^{-1} = \pi_n$ for each n, where π_n was defined to be the distribution of $(Y_1,\dots,Y_n)$. Thus, the sequences $\bigl(g_n(X'_n)\bigr)$ and (Y_n) have the same finite-dimensional distributions; in other words, their distributions agree on $\bigl\{\prod_{j=1}^N B_j \times \prod_{j=N+1}^\infty E'_j : N \in \mathbb{N}, B_j \in \mathcal{B}'_j \text{ for all } 1 \le j \le N\bigr\}$, a collection of cylindrical sets that generate the cylindrical σ-algebra $\mathcal{B}'$ of $\prod_{n=1}^\infty E'_n$. (By Kallenberg (1997, Lemma 1.2), $\mathcal{B}'$ is the Borel σ-algebra of $\prod_{n=1}^\infty E'_n$.) We conclude that $\bigl(g_1(X'_1), g_2(X'_2),\dots\bigr) \overset{d}{=} (Y_1, Y_2, \dots)$ as random sequences taking values in $\bigl(\prod_{n=1}^\infty E'_n, \mathcal{B}'\bigr)$.

Finally, we apply Kallenberg (1997, Corollary 5.11) with $T = \prod_{n=1}^\infty E_n$, $S = \prod_{n=1}^\infty E'_n$, $\eta = (X'_n)$, $\xi = (Y_n)$ and $f\colon T \to S$ given by $f(x_1, x_2, \dots) = \bigl(g_1(x_1), g_2(x_2), \dots\bigr)$; note that T, S are Polish spaces (e.g. Dudley, 2002, Theorem 2.5.7) and that f is Borel measurable. Having already shown that $f(\eta) \overset{d}{=} \xi$, we deduce from Kallenberg (1997, Corollary 5.11) that there exists $(\tilde{X}_n) \equiv \tilde{\eta} \overset{d}{=} \eta = (X'_n)$ satisfying $\bigl(g_n(\tilde{X}_n)\bigr) = f(\tilde{\eta}) = \xi = (Y_n)$ almost surely, as required. □

In the proofs of Proposition 6.16(a, c), we apply the concentration inequality below for sums of pseudo-Lipschitz functions of independent Gaussian random variables.

Lemma 7.12. *There exists a universal constant $C > 0$ such that the following holds for all $n \in \mathbb{N}$, $r \ge 2$ and $t \ge 0$: if $Z_1,\dots,Z_n \overset{\text{iid}}{\sim} N(0,1)$, $L \equiv (L_1,\dots,L_n) \in (0,\infty)^n$ and $f_i \in \mathrm{PL}_1(r, L_i)$ for $1 \le i \le n$, then*

$$\mathbb{P}\biggl(\biggl|\frac{1}{n}\sum_{i=1}^n \bigl\{f_i(Z_i) - \mathbb{E}\bigl(f_i(Z_i)\bigr)\bigr\}\biggr| \ge t\biggr) \le \exp\biggl(1 - \min\biggl\{\biggl(\frac{nt}{(Cr)^r\|L\|_2}\biggr)^2, \biggl(\frac{nt}{(Cr)^r\|L\|_\infty}\biggr)^{2/r}\biggr\}\biggr). \tag{7.6}$$

Proof. We first consider the case $n = 1$. For arbitrary $r \ge 2$ and $L > 0$, we may assume without loss of generality that $f \equiv f_1 \in \mathrm{PL}(r, L)$ satisfies $f(0) = 0$, so that $|f(x)| = |f(x) - f(0)| \le L(|x| + |x|^r) \le 2L(|x| \vee |x|^r)$ for all $x \in \mathbb{R}$. Thus, if $Z \sim N(0,1)$, then

$$\mathbb{P}(|f(Z)| \ge s) \le \mathbb{P}\bigl(|Z| \vee |Z|^r \ge s/(2L)\bigr) \le e^{-\frac{1}{2}\min\left\{\left(\frac{s}{2L}\right)^2, \left(\frac{s}{2L}\right)^{2/r}\right\}} \tag{7.7}$$

for all $s \geq 0$, and

$$\frac{\mathbb{E}\bigl(|f(Z)|\bigr)}{L} \leq \mathbb{E}(|Z| + |Z|^r) = \left(\sqrt{\frac{2}{\pi}} + \frac{2^{r/2}}{\sqrt{\pi}}\,\Gamma\left(\frac{r+1}{2}\right)\right) =: \upsilon_r$$

by direct computation. Now $\Gamma(x) < e^{1/(12x)}(x/e)^x\sqrt{2\pi/x}$ for all $x > 0$ by a non-asymptotic version of Stirling's formula; see for example Gordon (1994, Theorem 5) and Dümbgen et al. (2021, Lemma 10). Since $r \geq 2$, we have $(r+1)/e < r$ and $(\sqrt{2}-1)r^{r/2} \geq 2(\sqrt{2}-1) > 1/\sqrt{\pi}$. Therefore,

$$\frac{\upsilon_r}{2} \leq \frac{1}{\sqrt{2}}\left(\frac{1}{\sqrt{\pi}} + \left(\frac{r+1}{e}\right)^{r/2} e^{\frac{1}{6(r+1)}-\frac{1}{2}}\right) \leq \frac{1}{\sqrt{2}}\left(\frac{1}{\sqrt{\pi}} + r^{r/2}\right) < r^{r/2}. \tag{7.8}$$

Thus, for $t \geq L\upsilon_r$, we deduce from (7.7) and (7.8) that

$$\begin{aligned}\mathbb{P}\bigl\{\bigl|f(Z) - \mathbb{E}\bigl(f(Z)\bigr)\bigr| \geq t\bigr\} \leq \mathbb{P}\bigl\{|f(Z)| \geq t - \mathbb{E}\bigl(|f(Z)|\bigr)\bigr\} &\leq e^{-\frac{1}{2}\min\bigl\{\left(\frac{t}{2L}-\frac{\upsilon_r}{2}\right)^2,\,\left(\frac{t}{2L}-\frac{\upsilon_r}{2}\right)^{2/r}\bigr\}} \\ &\leq e^{\frac{1}{2}-\frac{1}{2}\left(\frac{t}{2L}-\frac{\upsilon_r}{2}\right)^{2/r}} \\ &\leq e^{\frac{1+r}{2}-\frac{1}{2}\left(\frac{t}{2L}\right)^{2/r}} \\ &\leq e^{1-\left(\frac{t}{2L}\right)^{2/r}\frac{1}{r+1}},\end{aligned} \tag{7.9}$$

where the third inequality follows from the fact that $a^{2/r} \leq |a-b|^{2/r} + b^{2/r}$ for $r \geq 2$ as above and any $a, b \geq 0$. Now (7.9) holds trivially for all $t \in [0, L\upsilon_r)$ since $1 - (r+1)^{-1}\{t/(2L)\}^{2/r} > 1 - (r+1)^{-1}(\upsilon_r/2)^{2/r} > 0$ by (7.8), so (7.6) holds with $C = 3$ when $n = 1$.

We now derive (7.6) for general $n \geq 2$ with the aid of Theorem 3.1 and Proposition A.3 in Kuchibhotla and Chakrabortty (2018); see also Theorem 1 and Corollary 2 in Bakhshizadeh et al. (2020). As in Sections 2 and 3 of Kuchibhotla and Chakrabortty (2018), we begin by defining $\vartheta_\beta \colon [0,\infty) \to [0,\infty)$ for each $\beta > 0$ by $\vartheta_\beta(x) := \exp(x^\beta) - 1$. Moreover, for $\beta, \lambda > 0$, let $\vartheta_{\beta,\lambda} \colon [0,\infty) \to [0,\infty)$ be the continuous, strictly increasing function with inverse given by $\vartheta_{\beta,\lambda}^{-1}(t) := \log^{1/2}(1+t) + \lambda\log^{1/\beta}(1+t)$ for $t \geq 0$. For a random variable X and a strictly increasing function $g \colon [0,\infty) \to [0,\infty)$ satisfying $g(0) = 0$, we write $\Xi_g(X) := \inf\bigl\{\theta > 0 : \mathbb{E}\bigl(g(|X|/\theta)\bigr) \leq 1\bigr\} \in [0,\infty]$, setting $\inf\emptyset = \infty$ by convention. Note that $\Xi_g(X)$ is precisely the g-Orlicz norm of X when g is convex, but that Ξ_g does not in general define a norm when g is not convex (for example when $g = \vartheta_\beta$ for $\beta \in (0,1)$, as in the proof below).

For arbitrary $n \geq 2$, $r \geq 2$ and $L \equiv (L_1,\ldots,L_n) \in (0,\infty)^n$, let $f_1,\ldots,f_n \in \mathrm{PL}(r, L_i)$ and $Z_1,\ldots,Z_n \overset{\mathrm{iid}}{\sim} N(0,1)$, and assume without loss of generality that $X_i := f_i(Z_i)$ satisfies $\mathbb{E}(X_i) = 0$ for all $1 \leq i \leq n$. Setting $\beta := 2/r \in [0,1]$ and $\theta_i := 2\{4(r+1)\}^{r/2}\,L_i$ for $1 \leq i \leq n$, we now integrate up the bound (7.9) to see that

$$\begin{aligned}\mathbb{E}\bigl(\vartheta_\beta(|X_i|/\theta_i)\bigr) = \int_0^\infty \mathbb{P}\bigl(\vartheta_\beta(|X_i|/\theta_i) \geq t\bigr)\,dt &= \int_0^\infty \mathbb{P}\bigl(|X_i| \geq \theta_i\vartheta_\beta^{-1}(t)\bigr)\,dt \\ &= \int_0^\infty \mathbb{P}\bigl(|X_i| \geq 2(r+1)^{r/2}L_i\,\{4\log(1+t)\}^{1/\beta}\bigr)\,dt \\ &\leq \int_0^\infty e(1+t)^{-4}\,dt = e/3 < 1,\end{aligned} \tag{7.10}$$

whence $\Xi_{\vartheta_\beta}(X_i) \leq \theta_i = 2\{4(r+1)\}^{r/2}\,L_i < \infty$. This shows that $X_1,\ldots,X_n$ are independent, centred *sub-Weibull* random variables of order $\beta = 2/r$, in the sense of Definition 2.2 in Kuchibhotla and Chakrabortty (2018). Then applying Kuchibhotla and Chakrabortty (2018, Theorem 3.1) with $a = (1/n,\ldots,1/n) \in \mathbb{R}^n$ and $b := \bigl(\Xi_{\vartheta_\beta}(X_1)/n,\ldots,\Xi_{\vartheta_\beta}(X_n)/n\bigr)$ in their notation, we deduce from (7.10) that

$$\Xi_{\vartheta_{\beta,\lambda_\beta}}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \leq 2eC_\beta\|b\|_2, \text{ where } \begin{cases} C_\beta := (2e^{2/e}/\beta)^{1/\beta}(128\pi)^{1/4}\,e^{3+\frac{1}{24}} \\ \lambda_\beta := (4^{1/\beta}/\sqrt{2})\,\|b\|_\infty/\|b\|_2. \end{cases} \tag{7.11}$$

It then follows from Proposition A.3 in Kuchibhotla and Chakrabortty (2018) that

$$\mathbb{P}\biggl(\biggl|\frac{1}{n}\sum_{i=1}^n X_i\biggr| \geq 4eC_\beta\|b\|_2\max\bigl(s^{1/2},\lambda_\beta s^{1/\beta}\bigr)\biggr) \leq e^{1-s}$$

for all $s \geq 0$, and hence that

$$\mathbb{P}\biggl(\biggl|\frac{1}{n}\sum_{i=1}^n X_i\biggr| \geq t\biggr) \leq \exp\biggl(1-\min\biggl\{\biggl(\frac{t}{4eC_\beta\|b\|_2}\biggr)^2,\biggl(\frac{t}{4eC'_\beta\|b\|_\infty}\biggr)^\beta\biggr\}\biggr)$$

for all $t \geq 0$, where $C'_\beta := (4^{1/\beta}/\sqrt{2})\,C_\beta$. Since $\beta = 2/r$ and $\Xi_{\vartheta_\beta}(X_i) \leq 2\{4(r+1)\}^{r/2}\,L_i$ for $1 \leq i \leq n$, we have

$$n\|b\|_p = \bigl\|\bigl(\Xi_{\vartheta_\beta}(X_1),\dots,\Xi_{\vartheta_\beta}(X_n)\bigr)\bigr\|_p \leq 2\{4(r+1)\}^{r/2}\,\|L\|_p$$

for $p \in \{2,\infty\}$. Moreover, $2\{4(r+1)\}^{r/2}\,C_\beta \leq 2\{4(r+1)\}^{r/2}\,C'_\beta \lesssim \{4e^{1/e}(r+1)\}^r$, so we can indeed find a suitable universal constant $C > 0$ in (7.6) such that the desired conclusion holds for all $n \in \mathbb{N}$, $r \geq 2$, $L \equiv (L_1,\dots,L_n) \in (0,\infty)^n$ and $t \geq 0$, as required. □

Remark 7.13. When $r > 2$, $f \in \mathrm{PL}_1(r)$ and $Z \sim N(0,1)$, the moment generating function of $f(Z)$ may not be finite anywhere except at 0 if $f(Z)$ has heavier tails than an exponential random variable (for example when $f(z) = \mathrm{sgn}(z)|z|^r$ for $z \in \mathbb{R}$). In these situations, the standard Chernoff method fails, which is why we apply different techniques that can handle general sub-Weibull random variables.

While we are primarily concerned with the case $r \geq 2$ in the proof of Proposition 6.16, there is an analogue of (7.6) when $r \in [1,2)$, namely

$$\mathbb{P}\biggl(\biggl|\frac{1}{n}\sum_{i=1}^n \bigl\{f_i(Z_i) - \mathbb{E}\bigl(f_i(Z_i)\bigr)\bigr\}\biggr| \geq t\biggr) \leq \exp\biggl(1-\min\biggl\{\biggl(\frac{nt}{C\|L\|_2}\biggr)^2,\biggl(\frac{nt}{C\|L\|_{\tilde{r}}}\biggr)^{2/r}\biggr\}\biggr), \tag{7.12}$$

where $C > 0$ is a suitable universal constant and $\tilde{r} := 2/(2-r) \in [2,\infty)$ is the Hölder conjugate of $2/r$. This can be proved using a Chernoff bound (e.g. Boucheron et al., 2013, Exercise 2.27), or alternatively using Kuchibhotla and Chakrabortty (2018, Theorem 3.1) once again, where we instead take $\beta := 2/r$, $C_\beta := 4e + 2(2\log 2)^{r/2}$ and $\lambda_\beta := (4^{1+1/\beta}C_\beta^{-1}e/\sqrt{2})\,\|b\|_{\tilde{r}}/\|b\|_2$ in (7.11).

The proof of Proposition 6.16(g) makes use of the following straightforward consequence of the definition of weak convergence.

Lemma 7.14. *On a Euclidean space E, if (μ_n) is a sequence of Borel probability measures that converges weakly to a Borel probability measure μ, then $\int_E g\,d\mu_n \to \int_E g\,d\mu$ for any bounded, Borel measurable $g\colon E \to \mathbb{R}$ that is continuous μ-almost everywhere (in the sense that the set of discontinuities of g has μ-measure 0).*

Proof. Writing $A \subseteq E$ for the set of discontinuities of g, we have $\mu(A) = 0$ by assumption. By Skorokhod's representation theorem (e.g. Kallenberg, 1997, Theorem 3.30), there exist random variables $X, X_1, X_2, \dots$ defined on a common probability space such that $X \sim \mu$, $X_n \sim \mu_n$ for all n and $X_n \to X$ almost surely. Then $g(X_n) \to g(X)$ almost surely on the event $\{X \in A^c\}$, which has probability $\mu(A^c) = 1$, so an application of the dominated (or bounded) convergence theorem shows that $\int_E g\,d\mu_n = \mathbb{E}\bigl(g(X_n)\bigr) \to \mathbb{E}\bigl(g(X)\bigr) = \int_E g\,d\mu$, as required. □

Remark 7.15. For each Lipschitz function $f_k\colon \mathbb{R}^2 \to \mathbb{R}$ in the AMP recursion (2.1), we assume in (A5) that there exists some f'_k that satisfies the hypotheses of Lemma 7.14 above with $\mu = \lambda \otimes \pi$; recall that λ denotes Lebesgue measure on $\mathbb{R}$ and the probability distribution π is as in (A1). To see why (A5) is a non-vacuous (albeit very mild) condition, consider Borel probability measures on $\mathbb{R}^D$ of the form $\mu = \lambda \otimes \nu$, where $D \geq 2$ and ν is some probability measure on $\mathbb{R}^{D-1}$. We will now give an example of a Lipschitz function $G\colon \mathbb{R}^D \to \mathbb{R}$ whose partial derivative $\frac{\partial G}{\partial x_1}$ cannot be extended beyond its domain of definition to a function $g\colon \mathbb{R}^D \to \mathbb{R}$ that is continuous μ-almost everywhere, for any μ of the above form.

Denote by $C \subseteq [0,1]$ the *fat Cantor set* (e.g. Aliprantis and Burkinshaw, 1998, pp. 140–141), which has the property that for all $x \in C$ and $\varepsilon > 0$, both $(x-\varepsilon, x+\varepsilon) \cap C$ and $(x-\varepsilon, x+\varepsilon) \cap C^c$ have positive Lebesgue measure. Then for any $f\colon \mathbb{R} \to \mathbb{R}$ with $f = \mathbb{1}_C$ Lebesgue almost everywhere, we have $\{f(u) : u \in (x-\varepsilon, x+\varepsilon)\} = \{0,1\}$ for all $x \in C$ and $\varepsilon > 0$, so f is discontinuous on C, which has Lebesgue measure $1/2 > 0$. Note that $F\colon x \mapsto \int_{-\infty}^{x} \mathbb{1}_C(t)\,dt$ is a Lipschitz function on $\mathbb{R}$ with $F'(x) = \mathbb{1}_C(x)$ for Lebesgue almost every $x \in \mathbb{R}$. Thus, for general $D \in \mathbb{N}$, the function $G\colon (x_1, \dots, x_D) \mapsto F(x_1)$ is Lipschitz on $\mathbb{R}^D$. Moreover, if $g\colon \mathbb{R}^D \to \mathbb{R}$ agrees with $\frac{\partial G}{\partial x_1}$ wherever the latter is defined, then g is discontinuous on $C \times \mathbb{R}^{D-1}$, which has strictly positive μ-measure when $\mu = \lambda \otimes \nu$ as above.

7.4 Wasserstein convergence and pseudo-Lipschitz functions

Throughout this subsection, we fix $D \in \mathbb{N}$ and $r \in [1,\infty)$, and write $\mathcal{P}(r) \equiv \mathcal{P}_D(r)$ for the set of probability measures P on $\mathbb{R}^D$ with $\int_{\mathbb{R}^D} \|x\|^r\,dP(x) < \infty$ (i.e. a finite r^{th} moment). For $P, Q \in \mathcal{P}(r)$, recall from Section 1.1 the definitions of $\widetilde{d}_r(P,Q)$ and the r-Wasserstein distance $d_r(P,Q)$.

The primary purpose of this subsection is to establish Theorem 7.17 and its probabilistic Corollary 7.21, which can be viewed as extensions of Villani (2003, Theorem 7.12). These show in particular that $\widetilde{d}_r$ and d_r are metrics on $\mathcal{P}(r)$ that induce the same topology (Remark 7.18), and also formalise the link between functions in $\mathrm{PL}_D(r)$ and convergence in d_r (or equivalently $\widetilde{d}_r$).

As a first step towards the proof of Theorem 7.17, it is helpful to establish the following.

Proposition 7.16. *There exists a countable set T_r' of bounded Lipschitz functions on $\mathbb{R}^D$ with the property that $\widetilde{d}_r(P,Q) = \sup_{\psi \in T_r'} \bigl|\int_{\mathbb{R}^D} \psi\,dP - \int_{\mathbb{R}^D} \psi\,dQ\bigr| \in [0,\infty)$ for all $P, Q \in \mathcal{P}(r)$.*

A key property of the set T_r' we construct is that for any $\psi \in \mathrm{PL}_D(r)$, there exists a sequence (ψ_ℓ) in T_r' that converges uniformly to ψ on compact subsets of $\mathbb{R}^D$. In subsequent proofs, we will write $Q(f)$ as shorthand for $\int_{\mathbb{R}^D} f\,dQ$ when Q is a signed Borel measure on $\mathbb{R}^D$ and $f\colon \mathbb{R}^D \to \mathbb{R}$ is a Q-integrable function.

Proof. For $N \in \mathbb{N}$, let $B_N \equiv \bar{B}_D(0,N) := \{x \in \mathbb{R}^D : \|x\| \le N\}$ and define $f_N(x) := (N - \|x\|) \vee 0 \wedge 1$ for $x \in \mathbb{R}^D$, so that f_N is 1-Lipschitz on $\mathbb{R}^D$, $f_N = 1$ on B_{N-1} and $f_N = 0$ on B_N^c. In the argument below (and in the proof of Theorem 7.17), we will use f_N as a substitute for the (discontinuous) indicator function $\mathbb{1}_{B_N}$ in several places. Note in particular that if $\tilde{g}\colon B_N \to \mathbb{R}$ is Lipschitz on B_N, then the function $g\colon \mathbb{R}^D \to \mathbb{R}$ defined by $g(x) := \tilde{g}(x) f_N(x)$ is Lipschitz and supported on the compact set B_N.

Recalling the definitions of $\widetilde{d}_r, \mathcal{P}(r)$ from (1.4) and writing $\widetilde{\mathrm{PL}}_D(r,1)$ for the set of all $\phi \in \mathrm{PL}_D(r,1)$ satisfying $\phi(0) = 0$, we see from (1.3) that

$$\widetilde{d}_r(P,Q) = \sup_{\phi \in \widetilde{\mathrm{PL}}_D(r,1)} |(P-Q)(\phi)| \le \sup_{\phi \in \widetilde{\mathrm{PL}}_D(r,1)} (P+Q)(|\phi|) \le \int_{\mathbb{R}^D} (\|x\| + \|x\|^r)\,d(P+Q)(x) < \infty \tag{7.13}$$

for all $P, Q \in \mathcal{P}(r)$. If $\phi \in \widetilde{\mathrm{PL}}_D(r,1)$, then

$$\bigl|\phi(x) - \phi(y)\bigr| \le \|x-y\|\bigl(1 + \|x\|^{r-1} + \|y\|^{r-1}\bigr) \le (1 + 2N^{r-1})\|x-y\|$$

for all $x, y \in B_N$, so $\phi|_{B_N}$ belongs to the set of $(1+2N^{r-1})$-Lipschitz functions $g\colon B_N \to \mathbb{R}$ satisfying $g(0) = 0$, which we denote by $\mathcal{G}_N$. Since B_N is compact and $\mathcal{G}_N$ is uniformly bounded and equicontinuous, $\mathcal{G}_N$ is therefore compact for the supremum norm on B_N by the Arzelà–Ascoli theorem (e.g. Dudley, 2002, Theorem 2.4.7). It is therefore totally bounded, so for each $m \in \mathbb{N}$, we can find a finite subset $\widetilde{\mathcal{H}}_{N,m} \subseteq \mathcal{G}_N$ such that for any $g \in \mathcal{G}_N$, there exists $\tilde{h} \in \widetilde{\mathcal{H}}_{N,m}$ with $\sup_{x \in B_N} |g(x) - \tilde{h}(x)| < 1/m$. Each $\tilde{h} \in \widetilde{\mathcal{H}}_{N,m}$ can be associated with a function $h\colon \mathbb{R}^D \to \mathbb{R}$ defined by $h(x) := \tilde{h}(x) f_N(x)$. By the reasoning in the previous paragraph, the collection $\mathcal{H}_{N,m}$ of all such h is a finite set of bounded Lipschitz functions supported on B_N.

Consequently, $T_r' := \bigcup_{N,m\in\mathbb{N}} \mathcal{H}_{N,m}$ is a countable set of bounded Lipschitz functions on $\mathbb{R}^D$, and we claim that this has the desired property that $\widetilde{d}_r(P,Q) = \sup_{\psi\in T_r'} |(P-Q)(\psi)|$ for any two probability measures $P,Q\in\mathcal{P}(r)$. Indeed, for fixed $P,Q\in\mathcal{P}(r)$, the function $\psi_r\colon x\mapsto \|x\|+\|x\|^r$ is integrable with respect to both P and Q on $\mathbb{R}^D$, so by the dominated convergence theorem, we have $P\bigl(\psi_r\mathbb{1}_{B_{N-1}^c}\bigr)\to 0$ and $Q\bigl(\psi_r\mathbb{1}_{B_{N-1}^c}\bigr)\to 0$ as $N\to\infty$. Thus, for an arbitrary $\varepsilon>0$, there exists a sufficiently large $N\equiv N_{\varepsilon,r}\in\mathbb{N}$ such that $P\bigl(\psi_r\mathbb{1}_{B_{N-1}^c}\bigr)<\varepsilon/4$ and $Q\bigl(\psi_r\mathbb{1}_{B_{N-1}^c}\bigr)<\varepsilon/4$. Choosing $m\equiv m_\varepsilon\in\mathbb{N}$ such that $1/m<\varepsilon/4$, we deduce from the previous paragraph that for any $\phi\in\widetilde{\mathrm{PL}}_D(r,1)$, there exists $\tilde{h}\in\widetilde{\mathcal{H}}_{N,m}$ such that $\sup_{x\in B_N}|\phi(x)-\tilde{h}(x)|<1/m<\varepsilon/4$. Letting h be the corresponding function in $\mathcal{H}_{N,m}\subseteq T_r'$, we have

$$\begin{aligned}|(P-Q)(\phi)| &\leq \bigl|(P-Q)\bigl(\phi(1-f_N)\bigr)\bigr| + |(P-Q)(\phi f_N - h)| + |(P-Q)(h)|\\ &\leq (P+Q)\bigl(|\phi|(1-f_N)\bigr) + (P+Q)(|\phi f_N - h|) + \sup\nolimits_{\psi\in T_r'}|(P-Q)(\psi)| \end{aligned}\tag{7.14}$$

by the triangle inequality. Since $\phi\in\widetilde{\mathrm{PL}}_D(r,1)$, we have $|\phi(x)|=|\phi(x)-\phi(0)|\leq\|x\|+\|x\|^r=\psi_r(x)$ for all $x\in\mathbb{R}^D$, whence

$$(P+Q)\bigl(|\phi|(1-f_N)\bigr)\leq (P+Q)\bigl(|\phi|\mathbb{1}_{B_{N-1}^c}\bigr)\leq (P+Q)\bigl(\psi_r\mathbb{1}_{B_{N-1}^c}\bigr)<\varepsilon/2$$

by our choice of N and the fact that $0\leq 1-f_N\leq\mathbb{1}_{B_{N-1}^c}$. Moreover,

$$(P+Q)(|\phi f_N - h|)\leq 2\sup\nolimits_{x\in B_N}|\phi(x)f_N(x)-h(x)|\leq 2\sup\nolimits_{x\in B_N}|\phi(x)-\tilde{h}(x)|<2/m<\varepsilon/2$$

by our choice of h, so it follows from (7.14) that $|(P-Q)(\phi)|<\varepsilon+\sup_{\psi\in T_r'}|(P-Q)(\psi)|$. Since this holds for every $\phi\in\widetilde{\mathrm{PL}}_D(r,1)$ and all $\varepsilon>0$, the result follows. □

Theorem 7.17. *Let $P\in\mathcal{P}_D(r)$ and let (P_n) be a sequence of probability measures in $\mathcal{P}_D(r)$. Then there exists a countable set $T_r\subseteq\mathrm{PL}_D(r)$ such that the following are equivalent:*

(i) $\int_{\mathbb{R}^D}\psi\,dP_n\to\int_{\mathbb{R}^D}\psi\,dP$ *for all* $\psi\in T_r$*;*

(ii) $\widetilde{d}_r(P_n,P)\to 0$*;*

(iii) $d_r(P_n,P)\to 0$*.*

A suitable set $T_r\subseteq\mathrm{PL}_D(r)$ can be constructed by enlarging the set T_r' of bounded Lipschitz functions defined in (the proof of) Proposition 7.16.

Remark 7.18. Using Theorem 7.17, we can verify that $\widetilde{d}_r$ is a metric on $\mathcal{P}(r)\equiv\mathcal{P}_D(r)$ that generates the same topology as d_r. Indeed, it is clear from (1.4) and (7.13) that $\widetilde{d}_r$ takes values in $[0,\infty)$ and satisfies the triangle inequality on $\mathcal{P}(r)$. In addition, if $P,Q\in\mathcal{P}(r)$ are such that $\widetilde{d}_r(P,Q)=0$, then by taking $P_n=Q$ for all n in (ii) above, we deduce that $d_r(P,Q)=0$. Since d_r is a metric on $\mathcal{P}(r)$ (e.g. Villani, 2003, Theorem 7.3), this yields $P=Q$, as required. In fact, $\bigl(\mathcal{P}(r),d_r\bigr)$ is a separable, complete metric space (e.g. Panaretos and Zemel, 2020, Theorem 2.2.7 and Proposition 2.2.8), so by the equivalence (ii) ⇔ (iii) in Theorem 7.17, the same is true of $\bigl(\mathcal{P}(r),\widetilde{d}_r\bigr)$.

Proof. (i) ⇒ (ii): As in the proof of Proposition 7.16, the function $f_N\colon x\mapsto (N-\|x\|)\vee 0\wedge 1$ once again serves as a Lipschitz surrogate for the indicator function $\mathbb{1}_{B_N}$ of $B_N\equiv\bar{B}_D(0,N)=\{x\in\mathbb{R}^D : \|x\|\leq N\}$ for each $N\in\mathbb{N}$ in the argument below; note that $f_N=1$ on B_{N-1}, $f_N=0$ on B_N^c and f_N is 1-Lipschitz on $\mathbb{R}^D$. In view of this and the fact that $\psi_r\colon x\mapsto\|x\|+\|x\|^r$ belongs to $\mathrm{PL}_D(r)$, the function $\psi_r(1-f_N)$ also lies in $\mathrm{PL}_D(r)$ for every $N\in\mathbb{N}$.

Let $\widetilde{\mathcal{H}}_{N,m}$ and $\mathcal{H}_{N,m}$ be the finite sets constructed in the proof of Proposition 7.16 for each $N,m\in\mathbb{N}$, and let $T_r' := \bigcup_{N,m\in\mathbb{N}}\mathcal{H}_{N,m}$. Since T_r' is a set of bounded Lipschitz functions, we certainly have $T_r'\subseteq\mathrm{PL}_D(r)$. We claim that $T_r := T_r'\cup\{\psi_r(1-f_N) : N\in\mathbb{N}\}$ is a countable subset of $\mathrm{PL}_D(r)$ with the required property. To see this, suppose that (i) holds for this set T_r, i.e. that $P_n(\psi)\to P(\psi)$ for all $\psi\in T_r$. As noted in (7.13), we have $\widetilde{d}_r(P_n,P)=\sup_{\phi\in\widetilde{\mathrm{PL}}_D(r,1)}|(P_n-P)(\phi)|$ for all n, where $\widetilde{\mathrm{PL}}_D(r,1)$

denotes the set of all $\phi \in \mathrm{PL}_D(r,1)$ satisfying $\phi(0)=0$, so it suffices to show that the latter quantity converges to 0.

We will consider a decomposition (7.15) similar to (7.14) in the proof of Proposition 7.16, taking particular care in this instance to ensure that the subsequent bounds hold uniformly over $\phi \in \widetilde{\mathrm{PL}}_D(r,1)$. Observe that since $\psi_r(1-f_N) \to 0$ pointwise on $\mathbb{R}^D$ as $N \to \infty$, and $\psi_r(1-f_N)$ is dominated by the P-integrable function ψ_r on $\mathbb{R}^D$ for each N, we have $P\bigl(\psi_r(1-f_N)\bigr) \to 0$ as $N \to \infty$ by the dominated convergence theorem. Thus, for an arbitrary $\varepsilon > 0$, there exists a sufficiently large $N \equiv N_{\varepsilon,r} \in \mathbb{N}$ such that $P\bigl(\psi_r(1-f_N)\bigr) < \varepsilon/4$, and we also fix $m \equiv m_\varepsilon \in \mathbb{N}$ such that $1/m < \varepsilon/4$. With this choice of N and m, it follows from the defining property of $\widetilde{\mathcal{H}}_{N,m}$ that for any $\phi \in \widetilde{\mathrm{PL}}_D(r,1)$, there exists $\tilde{h}_\phi \in \widetilde{\mathcal{H}}_{N,m}$ such that $\sup_{x\in B_N}|\phi(x)-\tilde{h}_\phi(x)| < 1/m < \varepsilon/4$. Letting h_ϕ be the corresponding function in $\mathcal{H}_{N,m}$ as above, we have

$$\begin{aligned}|(P_n-P)(\phi)| &\le \bigl|(P_n-P)\bigl(\phi(1-f_N)\bigr)\bigr| + |(P_n-P)(\phi f_N - h_\phi)| + |(P_n-P)(h_\phi)| \\ &\le (P_n+P)\bigl(|\phi|(1-f_N)\bigr) + (P_n+P)(|\phi f_N - h_\phi|) + \max_{\psi\in\mathcal{H}_{N,m}} |(P_n-P)(\psi)| \end{aligned} \tag{7.15}$$

by the triangle inequality. Now for every $\phi \in \widetilde{\mathrm{PL}}_D(r,1)$, we have $|\phi(x)| = |\phi(x)-\phi(0)| \le \|x\| + \|x\|^r = \psi_r(x)$ for all $x \in \mathbb{R}^D$. Since $P_n\bigl(\psi_r(1-f_N)\bigr) \to P\bigl(\psi_r(1-f_N)\bigr)$ as $n \to \infty$ by assumption, this implies that

$$\limsup_{n\to\infty} \sup_{\phi\in\widetilde{\mathrm{PL}}_D(r,1)} (P_n+P)\bigl(|\phi|(1-f_N)\bigr) \le \limsup_{n\to\infty} (P_n+P)\bigl(\psi_r(1-f_N)\bigr) = 2P\bigl(\psi_r(1-f_N)\bigr) < \varepsilon/2. \tag{7.16}$$

Moreover, for any $\phi \in \widetilde{\mathrm{PL}}_D(r,1)$, the functions ϕf_N and h_ϕ are both supported on B_N, and $|\phi f_N - h_\phi| = |\phi - \tilde{h}_\phi|\, f_N \le |\phi - \tilde{h}_\phi| < \varepsilon/4$ on B_N, so

$$\limsup_{n\to\infty} \sup_{\phi\in\widetilde{\mathrm{PL}}_D(r,1)} (P_n+P)(|\phi f_N - h_\phi|) \le 2 \sup_{\phi\in\widetilde{\mathrm{PL}}_D(r,1)} \sup_{x\in B_N} |\phi(x)f_N(x) - h_\phi(x)| < \varepsilon/2. \tag{7.17}$$

Finally, since $\mathcal{H}_{N,m}$ is finite and $P_n(\psi) \to P(\psi)$ for all $\psi \in \mathcal{H}_{N,m} \subseteq \mathcal{T}_r$ by assumption, we have $\max_{\psi\in\mathcal{H}_{N,m}}|(P_n-P)(\psi)| \to 0$. Combining this with (7.15), (7.16) and (7.17), we conclude that

$$\limsup_{n\to\infty} \tilde{d}_r(P_n,P) = \limsup_{n\to\infty} \sup_{\phi\in\widetilde{\mathrm{PL}}_D(r,1)} |(P_n-P)(\phi)| < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, the desired conclusion follows.

(ii) ⇒ (iii): Suppose that $\tilde{d}_r(P_n,P) \to 0$ and let $\psi\colon \mathbb{R}^D \to \mathbb{R}$ be a (bounded) L-Lipschitz function, for some $L > 0$. Then $\tilde{\psi}(\cdot) := \psi(\cdot)/L \in \mathrm{PL}_D(r,1)$, so $P_n(\psi) = LP_n(\tilde{\psi}) \to LP(\tilde{\psi}) = P(\psi)$. Hence $P_n \xrightarrow{d} P$. Moreover, the function $x \mapsto \|x\|^r$ belongs to $\mathrm{PL}(r,(r/2)\vee 1)$ since by Lemma 7.20 below,

$$\bigl|\|x\|^r - \|y\|^r\bigr| \le \frac{r\vee 2}{2}\bigl|\|x\| - \|y\|\bigr|\bigl(\|x\|^{r-1} + \|y\|^{r-1}\bigr) \le \frac{r\vee 2}{2}\|x-y\|\bigl(\|x\|^{r-1} + \|y\|^{r-1}\bigr) \tag{7.18}$$

for all $x,y \in \mathbb{R}^D$, so $\int_{\mathbb{R}^D}\|x\|^r\,dP_n(x) \to \int_{\mathbb{R}^D}\|x\|^r\,dP(x)$. We conclude that $d_r(P_n,P) \to 0$.

(iii) ⇒ (i): We will show here that if (iii) holds, then $P_n(\psi) \to P(\psi)$ for all $\psi \in \mathrm{PL}_D(r)$. Indeed, suppose that $P_n \xrightarrow{d} P$ and $\int_{\mathbb{R}^D}\|x\|^r\,dP_n(x) \to \int_{\mathbb{R}^D}\|x\|^r\,dP(x)$. Now for $L > 0$ and any $\psi \in \widetilde{\mathrm{PL}}_D(r,L)$, we have $|\psi(x)| \le L\|x\|(1+\|x\|^{r-1}) \le 2L(1+\|x\|^r)$ for all $x \in \mathbb{R}^D$. Thus, since ψ is continuous on $\mathbb{R}^D$ and $x \mapsto |\psi(x)|/(1+\|x\|^r)$ is bounded on $\mathbb{R}^D$, it follows from (iii) and Dümbgen et al. (2011, Lemma 4.5) that $P_n(\psi) \to P(\psi)$. □

Remark 7.19. The proof of the implication (i) ⇒ (ii) in Theorem 7.17 is similar to the argument in Dudley (2002) showing that (b) implies (c) in his Theorem 11.3.3, where it is established that the bounded Lipschitz metric induces the topology of weak convergence (of probability measures on a separable metric space).

To obtain a sharp pseudo-Lipschitz constant for $x \mapsto \|x\|^r$ in (7.18) above, we apply the following elementary inequality.

Lemma 7.20. *If $a, b \geq 0$ and $r \geq 1$, then $|a^r - b^r| \leq \max(1, r/2)\, |a - b|\, (a^{r-1} + b^{r-1})$.*

Proof. Suppose without loss of generality that $0 \leq b \leq a$. If $r \geq 2$, then $t \mapsto rt^{r-1}$ is convex on $[0, \infty)$, so

$$a^r - b^r = \int_a^b rt^{r-1}\, dt \leq \int_a^b r\left(\frac{t-b}{a-b}\, a^{r-1} + \frac{a-t}{a-b}\, b^{r-1}\right) dt = \frac{r}{2}(a-b)(a^{r-1} + b^{r-1}).$$

If $r \in [1, 2]$, then $0 \leq (ab)^{r-1}(a^{2-r} - b^{2-r}) = ab^{r-1} - ba^{r-1}$, so $a^r - b^r \leq (a-b)(a^{r-1} + b^{r-1})$. □

When we have a sequence of possibly random probability measures $P_n \equiv P_n(\omega)$ on $\mathbb{R}^D$, we can apply the deterministic Theorem 7.17 to obtain Corollary 7.21 below, in which we equip $\mathcal{P}(r)$ with the Borel σ-algebra $\mathcal{B}_r \equiv \mathcal{B}\big(\mathcal{P}(r)\big)$ associated with the d_r (or equivalently the $\widetilde{d}_r$) metric. Note that $\widetilde{d}_r(P_n, P)$ is measurable (i.e. a bona fide random variable) for each n by Proposition 7.16. The measurability of $d_r(P_n, P)$ is guaranteed by Villani (2009, Corollary 5.22); see also Panaretos and Zemel (2020, Lemma 2.4.6).

Corollary 7.21. *Fix $P \in \mathcal{P}(r) \equiv \mathcal{P}_D(r)$ and let (P_n) be a sequence of random elements $P_n\colon \Omega \to \mathcal{P}(r)$.*

(a) Then the following are equivalent:

(i) $\int_{\mathbb{R}^D} \psi\, dP_n \overset{a.s.}{\to} \int_{\mathbb{R}^D} \psi\, dP$ for every $\psi \in \mathrm{PL}_D(r)$;

(ii) $\widetilde{d}_r(P_n, P) \overset{a.s.}{\to} 0$;

(iii) $d_r(P_n, P) \overset{a.s.}{\to} 0$.

(b) The same equivalences hold if the mode of convergence in (i)–(iii) is instead taken to be either convergence in probability or complete convergence.

Thus, to establish the seemingly stronger conclusions in (ii) and (iii) for a random sequence of distributions P_n, a putative limit $P \in \mathcal{P}_D(r)$ and any of the above modes of stochastic convergence, it is sufficient (and sometimes more convenient) to show that the appropriate version of (i) holds for each $\psi \in \mathrm{PL}_D(r)$ in turn. This is the approach we take in the proofs of the master theorems for symmetric AMP (Theorems 2.1 and 2.3).

Proof. (a) The implications (ii) ⇒ (iii) ⇒ (i) are immediate from Theorem 7.17. As for (i) ⇒ (ii), note that for each $\psi \in \mathrm{PL}_D(r)$ in (i), the event $\Omega(\psi)$ of probability 1 on which $\int_{\mathbb{R}^D} \psi\, dP_n \to \int_{\mathbb{R}^D} \psi\, dP$ may depend (a priori) on ψ. The key point is that under (i), Theorem 7.17 ensures that this convergence is actually uniform over $\mathrm{PL}_D(r, 1)$ on a *countable* intersection of such events $\Omega(\psi)$. More precisely, letting $T_r \subseteq \mathrm{PL}_D(r)$ be as in Theorem 7.17, we see that $\bigcap_{\psi \in T_r} \Omega(\psi)$ is an event of probability 1 on which (ii) and (iii) hold.

(b) *Convergence in probability*:

(i) ⇒ (ii): First, we prove that if $P_n(\psi) \overset{p}{\to} P(\psi)$ for each $\psi \in \mathrm{PL}_D(r)$, then $\widetilde{d}_r(P_n, P) \overset{p}{\to} 0$, or equivalently that every subsequence of $\big(\widetilde{d}_r(P_n, P) : n \in \mathbb{N}\big)$ has a further subsequence that converges almost surely to 0. It suffices to show that for any subsequence $(Q_k) \equiv (P_{n_k})$, there is a further subsequence (Q_{k_ℓ}) such that with probability 1, we have $Q_{k_\ell}(\psi) \to P(\psi)$ for all $\psi \in T_r \subseteq \mathrm{PL}_D(r)$; indeed, the desired conclusion then follows directly from (a). To this end, enumerate the elements of the countable set T_r as $\psi_1, \psi_2, \dots$ and apply a diagonal argument: since $Q_k(\psi_1) \overset{p}{\to} P(\psi_1)$, we can extract a subsequence $(Q_{k_{1,\ell}})$ of (Q_k) such that $Q_{k_{1,\ell}}(\psi_1) \overset{a.s.}{\to} P(\psi_1)$ as $\ell \to \infty$. Continuing inductively, we see that for each $J \in \mathbb{N}$, there exist a subsequence $(Q_{k_{J,\ell}})$ of $(Q_{k_{J-1,\ell}})$ and an event of probability 1 on which $Q_{k_{J,\ell}}(\psi_j) \to P(\psi_j)$ as $\ell \to \infty$ for all $1 \leq j \leq J$. Finally, let $Q_{k_\ell} := Q_{k_{\ell,\ell}}$ for $\ell \in \mathbb{N}$, and observe that with probability 1, we have $Q_{k_{J,\ell}}(\psi_j) \to P(\psi_j)$ as $\ell \to \infty$ for all $j \in \mathbb{N}$, as required.

(ii) $\Rightarrow$ (iii) $\Rightarrow$ (i): As above, we can argue along subsequences of (P_n) and then appeal directly to the corresponding implications in (a).

Complete convergence:

(i) $\Rightarrow$ (ii): Suppose that $P_n(\psi) \overset{c}{\to} P(\psi)$ for every $\psi \in \mathrm{PL}_D(r)$. In view of Definition 1.1 of complete convergence, it suffices to show that if (β_n) is any sequence of random variables with $\beta_n \overset{d}{=} \widetilde{d}_r(P_n, P)$ for each n, then $\beta_n \overset{a.s.}{\to} 0$. For any such sequence (β_n), we first seek to construct a sequence $\bigl(\tilde{P}_n\bigr)$ of random elements $\tilde{P}_n \colon \Omega \to \bigl(\mathcal{P}(r), \widetilde{d}_r\bigr)$ such that $\tilde{P}_n \overset{d}{=} P_n$ on $\bigl(\mathcal{P}(r), \mathcal{B}_r\bigr)$ for each n and $\bigl(\widetilde{d}_r(\tilde{P}_n, P) : n \in \mathbb{N}\bigr) = (\beta_n : n \in \mathbb{N})$ almost surely as random sequences. Since $\bigl(\mathcal{P}(r), \widetilde{d}_r\bigr)$ is a Polish space, a suitable $\bigl(\tilde{P}_n\bigr)$ can be obtained by applying Lemma 7.10, where for each n, we take $g_n \colon \bigl(\mathcal{P}(r), \widetilde{d}_r\bigr) \to \mathbb{R}$ to be the 1-Lipschitz (and hence Borel measurable) function $Q \mapsto \widetilde{d}_r(P, Q)$.

For each $\psi \in \mathrm{PL}_D(r)$, we see from the definition of $\widetilde{d}_r$ in (1.4) that $Q \mapsto Q(\psi) = \int_{\mathbb{R}^D} \psi \, dQ$ is also a 1-Lipschitz (and hence Borel measurable) function from $\bigl(\mathcal{P}(r), \widetilde{d}_r\bigr)$ to $\mathbb{R}$, so $\tilde{P}_n(\psi) \colon \Omega \to \mathbb{R}$ is measurable (i.e. a random variable). Now $\tilde{P}_n \overset{d}{=} P_n$ for each n by construction, so for every $\psi \in \mathrm{PL}_D(r)$, it follows that $\tilde{P}_n(\psi) \overset{d}{=} \tilde{P}_n(\psi)$ for each n and hence that $\tilde{P}_n(\psi) \overset{a.s.}{\to} P(\psi)$. Thus, by the implication (i) $\Rightarrow$ (ii) in (a) above, we conclude that $\beta_n = \widetilde{d}_r(\tilde{P}_n, P) \to 0$ almost surely, as required.

(ii) $\Rightarrow$ (iii) $\Rightarrow$ (i): To establish these remaining implications, observe that it suffices to show the following: if $F_n, G_n \colon \mathcal{P}(r) \to \mathbb{R}$ are Borel measurable functions for which it is known from (a) that $F_n(P_n) \overset{a.s.}{\to} 0$ implies $G_n(P_n) \overset{a.s.}{\to} 0$, then $F_n(P_n) \overset{c}{\to} 0$ implies $G_n(P_n) \overset{c}{\to} 0$. To prove this, we can proceed as in the argument for (i) $\Rightarrow$ (ii): given any random sequence (β_n) such that $\beta_n \overset{d}{=} G_n(P_n)$ for each n, Lemma 7.10 yields a sequence $(\tilde{P}_n)$ of random elements $\tilde{P}_n \colon \Omega \to \mathcal{P}(r)$ such that $F_n(\tilde{P}_n) \overset{d}{=} F_n(P_n)$ and $\beta_n = G_n(\tilde{P}_n)$ almost surely for each n. Then $F_n(\tilde{P}_n) \overset{a.s.}{\to} 0$, so (a) implies that $\beta_n = G_n(\tilde{P}_n) \to 0$ almost surely. This completes the proof. □

We conclude this subsection with some straightforward results on pseudo-Lipschitz functions.

Lemma 7.22. *For $D \in \mathbb{N}$, if $f \in \mathrm{PL}_D(r)$ and $g \in \mathrm{PL}_D(s)$ for some $r, s \geq 1$, then $fg \in \mathrm{PL}_D(r+s)$ and $|f|^p \in \mathrm{PL}_D(pr)$ for all $p \geq 1$.*

Proof. There exists $L > 0$ such that $f \in \mathrm{PL}_D(r, L)$ and $g \in \mathrm{PL}_D(s, L)$. Letting $L' := L \vee |f(0)| \vee |g(0)|$, we have

$$\begin{aligned} |f(x)| &\leq |f(0)| + |f(x) - f(0)| \leq L'(1 + \|x\| + \|x\|^r) \leq 2L'(1 + \|x\|^r) \\ |g(x)| &\leq |g(0)| + |g(x) - g(0)| \;\leq L'(1 + \|x\| + \|x\|^s) \leq 2L'(1 + \|x\|^s) \end{aligned} \tag{7.19}$$

for all $x \in \mathbb{R}^D$. Therefore, fixing arbitrary $x, y \in \mathbb{R}^D$ and setting $a := \|x\| \vee \|y\|$, we see that

$$\begin{aligned} &|f(x)g(x) - f(y)g(y)| \\ &\quad\leq |f(x)|\,|g(x) - g(y)| + |g(x)|\,|f(x) - f(y)| \\ &\quad\leq 2L'L\,\|x - y\|\,\bigl\{\bigl(1 + \|x\|^r\bigr)\bigl(1 + \|x\|^{s-1} + \|y\|^{s-1}\bigr) + \bigl(1 + \|x\|^s\bigr)\bigl(1 + \|x\|^{r-1} + \|y\|^{r-1}\bigr)\bigr\} \\ &\quad\leq 2L'L\,\|x - y\|\,(2 + 2a^{r-1} + a^r + 2a^{s-1} + a^s + 4a^{r+s-1}) \\ &\quad\leq 20L'L\,\|x - y\|\,(1 + a^{r+s-1}) \\ &\quad\leq 20L'L\,\|x - y\|\,(1 + \|x\|^{r+s-1} + \|y\|^{r+s-1}). \end{aligned}$$

This shows that $fg \in \mathrm{PL}_D(r+s)$. For $p \geq 1$, we have $(a+b)^{p-1} \leq (1 \vee 2^{p-2})(a^{p-1} + b^{p-1})$ for $a, b \geq 0$, and it follows from Lemma 7.20 and (7.19) that

$$\begin{aligned} \bigl||f(x)|^p - |f(y)|^p\bigr| &\leq \frac{p \vee 2}{2}\,|f(x) - f(y)|\,\bigl(|f(x)|^{p-1} + |f(y)|^{p-1}\bigr) \\ &\leq \frac{L(2L')^{p-1}(p \vee 2)}{2}\,\|x - y\|\,(1 + \|x\|^r + \|y\|^r)\bigl((1 + \|x\|^r)^{p-1} + (1 + \|y\|^r)^{p-1}\bigr) \\ &\lesssim_p L(L')^{p-1}\,\|x - y\|\,(1 + \|x\|^r + \|y\|^r)\bigl(1 + \|x\|^{(p-1)r} + \|y\|^{(p-1)r}\bigr) \\ &\lesssim_p L(L')^{p-1}\,\|x - y\|\,(1 + \|x\|^{pr} + \|y\|^{pr}) \end{aligned}$$

for all $x, y \in \mathbb{R}^D$. Thus, $|f|^p \in \mathrm{PL}_D(pr)$, as required. □

Lemma 7.23. *Let $\psi \in \mathrm{PL}_{D+1}(r, L)$ for some $D \in \mathbb{N}$, $r \geq 1$ and $L > 0$. Fix $c \equiv (c_1, \ldots, c_D) \in \mathbb{R}^D$ and $\tau > 0$.*

(a) For fixed $x \equiv (x_1, \ldots, x_D) \in \mathbb{R}^D$, define $\psi_x \colon \mathbb{R} \to \mathbb{R}$ by $\psi_x(z) := \psi\bigl(x_1, \ldots, x_D, \sum_{\ell=1}^D c_\ell\, x_\ell + \tau z\bigr)$. Then $\psi_x \in \mathrm{PL}_1(r, L_{\|x\|,\tau})$, where $L_{a,\tau} := L\tau \max\{1 + (2 \vee 2^{r-1})(1 + \|c\|)^{r-1} a^{r-1}, (1 \vee 2^{r-2})\tau^{r-1}\}$ for $a \geq 0$.

(b) Let $Z \sim N(0,1)$ and define $\Psi \colon \mathbb{R}^D \to \mathbb{R}$ by $\Psi(x_1, \ldots, x_D) := \mathbb{E}\bigl\{\psi\bigl(x_1, \ldots, x_D, \sum_{\ell=1}^D c_\ell\, x_\ell + \tau Z\bigr)\bigr\}$. Then $\Psi \in \mathrm{PL}_D(r, L_\tau)$, where $L_\tau := L(1 + \|c\|) \max\{1 + (2 \vee 2^{r-1})\,\mathbb{E}(|\tau Z|^{r-1}), (1 \vee 2^{r-2})(1 + \|c\|)^{r-1}\}$.

Proof. For $x \equiv (x_1, \ldots, x_D) \in \mathbb{R}^D$ and $z \in \mathbb{R}$, note first that

$$\begin{aligned}
\bigl\|\bigl(x_1, \ldots, x_D, \textstyle\sum_{\ell=1}^D c_\ell\, x_\ell + \tau z\bigr)\bigr\|^{r-1} &\leq \bigl(\|x\| + |\textstyle\sum_{\ell=1}^D c_\ell\, x_\ell| + \tau|z|\bigr)^{r-1} \\
&\leq \{(1 + \|c\|)\|x\| + \tau|z|\}^{r-1} && (7.20) \\
&\leq (1 \vee 2^{r-2})\bigl\{(1 + \|c\|)^{r-1}\|x\|^{r-1} + \tau^{r-1}|z|^{r-1}\bigr\}, && (7.21)
\end{aligned}$$

where the three bounds above are obtained using the triangle inequality, the Cauchy–Schwarz inequality and the fact that $(a + b)^{r-1} \leq (1 \vee 2^{r-2})(a^{r-1} + b^{r-1})$ for $a, b \geq 0$.

(a) For $z, z' \in \mathbb{R}$, we have

$$\begin{aligned}
&|\psi_x(z) - \psi_x(z')| \\
&\quad= \bigl|\psi\bigl(x_1, \ldots, x_D, \textstyle\sum_{\ell=1}^D c_\ell\, x_\ell + \tau z\bigr) - \psi\bigl(x_1, \ldots, x_D, \sum_{\ell=1}^D c_\ell\, x_\ell + \tau z'\bigr)\bigr| \\
&\quad\leq L\tau|z - z'|\,\bigl\{1 + 2(1 \vee 2^{r-2})(1 + \|c\|)^{r-1}\|x\|^{r-1} + (1 \vee 2^{r-2})\,\tau^{r-1}\bigl(|z|^{r-1} + |z'|^{r-1}\bigr)\bigr\} \\
&\quad\leq L_{\|x\|,\tau}\,|z - z'|\,\bigl(1 + |z|^{r-1} + |z'|^{r-1}\bigr),
\end{aligned}$$

where the first bound follows from (7.21) and the fact that $\psi \in \mathrm{PL}_{D+1}(r, L)$.

(b) For $x, y \in \mathbb{R}^D$, we have

$$\begin{aligned}
&|\Psi(x) - \Psi(y)| \\
&\quad\leq \mathbb{E}\bigl\{\bigl|\psi\bigl(x_1, \ldots, x_D, \textstyle\sum_{\ell=1}^D c_\ell\, x_\ell + \tau Z\bigr) - \psi\bigl(y_1, \ldots, y_D, \sum_{\ell=1}^D c_\ell\, y_\ell + \tau Z\bigr)\bigr|\bigr\} \\
&\quad\leq L(1 + \|c\|)\|x - y\|\,\bigl\{1 + 2(1 \vee 2^{r-2})\,\mathbb{E}(|\tau Z|^{r-1}) + (1 \vee 2^{r-2})(1 + \|c\|)^{r-1}\bigl(\|x\|^{r-1} + \|y\|^{r-1}\bigr)\bigr\} \\
&\quad\leq L_\tau\|x - y\|\,\bigl(1 + \|x\|^{r-1} + \|y\|^{r-1}\bigr),
\end{aligned}$$

where the second bound again follows from (7.20), (7.21) and the fact that $\psi \in \mathrm{PL}_{D+1}(r, L)$. □

Lemma 7.24. *Suppose that $\psi \in \mathrm{PL}_D(r, L)$ for some $D \in \mathbb{N}$, $r \in [2, \infty)$ and $L > 0$. Then for any $n \in \mathbb{N}$ and vectors $x^\ell \equiv (x_1^\ell, \ldots, x_n^\ell)$ and $y^\ell \equiv (y_1^\ell, \ldots, y_n^\ell)$ for $1 \leq \ell \leq D$, we have*

$$\frac{1}{n}\sum_{i=1}^n |\psi(x_i^1, \ldots, x_i^D) - \psi(y_i^1, \ldots, y_i^D)| \leq L D^{\frac{r}{2}-1}\left(\sum_{\ell=1}^D \|x^\ell - y^\ell\|_{n,r}^r\right)^{1/r}\left(1 + \sum_{\ell=1}^D \bigl(\|x^\ell\|_{n,r}^{r-1} + \|y^\ell\|_{n,r}^{r-1}\bigr)\right).$$

Proof. For $1 \leq i \leq n$, define $X^{(i)} := (x_i^1, \ldots, x_i^D)$ and $Y^{(i)} := (y_i^1, \ldots, y_i^D)$, and let $r' := r/(r-1) \in (1, 2]$ be the Hölder conjugate of r, so that $1/r + 1/r' = 1$. Then since $\psi \in \mathrm{PL}_D(r, L)$, an application of Hölder's inequality yields the bound

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n |\psi(x_i^1, \ldots, x_i^D) - \psi(y_i^1, \ldots, y_i^D)| &= \frac{1}{n}\sum_{i=1}^n |\psi(X^{(i)}) - \psi(Y^{(i)})| \\
&\leq \frac{1}{n}\sum_{i=1}^n L\|X^{(i)} - Y^{(i)}\|\,\bigl(1 + \|X^{(i)}\|^{r-1} + \|Y^{(i)}\|^{r-1}\bigr) \\
&\leq L\left(\frac{1}{n}\sum_{i=1}^n \|X^{(i)} - Y^{(i)}\|^r\right)^{1/r}\left(\frac{1}{n}\sum_{i=1}^n \bigl(1 + \|X^{(i)}\|^{r-1} + \|Y^{(i)}\|^{r-1}\bigr)^{r'}\right)^{1/r'}. \qquad (7.22)
\end{aligned}$$

Since $\|\cdot\| \equiv \|\cdot\|_2 \leq D^{\frac{1}{2}-\frac{1}{r}}\|\cdot\|_r$ on $\mathbb{R}^D$, we see that

$$\frac{1}{n}\sum_{i=1}^n \|X^{(i)} - Y^{(i)}\|^r \leq \frac{D^{\frac{r}{2}-1}}{n}\sum_{i=1}^n\sum_{\ell=1}^D |x_i^\ell - y_i^\ell|^r = D^{\frac{r}{2}-1}\sum_{\ell=1}^D \|x^\ell - y^\ell\|_{n,r}^r. \tag{7.23}$$

In addition, by applying the triangle inequality for $\|\cdot\|_{n,r'}$ and arguing as in (7.23), we have

$$\begin{aligned}\biggl(\frac{1}{n}\sum_{i=1}^n \bigl(1 + \|X^{(i)}\|^{r-1} + \|Y^{(i)}\|^{r-1}\bigr)^{r'}\biggr)^{1/r'} &\leq 1 + \biggl(\frac{1}{n}\sum_{i=1}^n \|X^{(i)}\|^r\biggr)^{1/r'} + \biggl(\frac{1}{n}\sum_{i=1}^n \|Y^{(i)}\|^r\biggr)^{1/r'} \\ &\leq 1 + \biggl(D^{\frac{r}{2}-1}\sum_{\ell=1}^D \|x^\ell\|_{n,r}^r\biggr)^{\frac{r-1}{r}} + \biggl(D^{\frac{r}{2}-1}\sum_{\ell=1}^D \|y^\ell\|_{n,r}^r\biggr)^{\frac{r-1}{r}} \\ &\leq 1 + (D^{\frac{r}{2}-1})^{\frac{r-1}{r}}\sum_{\ell=1}^D \bigl(\|x^\ell\|_{n,r}^{r-1} + \|y^\ell\|_{n,r}^{r-1}\bigr),\end{aligned} \tag{7.24}$$

where the final bound follows since $\|\cdot\|_r \leq \|\cdot\|_{r-1}$ on $\mathbb{R}^D$. Combining (7.22)–(7.24) yields the desired conclusion. □

References

Advani, M. S., Saxe, A. M. and Sompolinsky, H. (2020). High-dimensional dynamics of generalization error in neural networks. *Neural Netw.*, **132**, 428–446.

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models.* Wiley, New Jersey.

Albert, A. and Anderson J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, **71**, 1–10.

Aliprantis, C. D. and Burkinshaw, O. (1998). *Principles of Real Analysis*, 3rd edition. Academic Press, San Diego.

Alon, N., Krivelevich, M. and Sudakov, B. (1998). Finding a large hidden clique in a random graph. *Random Struct. Algorithms*, **13**, 457–466.

Anderson, G., Guionnet, A. and Zeitouni, O. (2010). *An Introduction to Random Matrices.* Cambridge University Press, Cambridge.

Aubin, B., Maillard, A., Barbier, J., Krzakala, F., Macris, N. and Zdeborová, L. (2019). The committee machine: computational to statistical gaps in learning a two-layers neural network. *J. Stat. Mech. Theory Exp.*, 124023.

Aubin, B., Loureiro, B., Maillard, A., Krzakala, F., Zdeborová, L. (2020). The spiked matrix model with generative priors. *IEEE Trans. Inf. Theory*, **67**, 1156–1181.

Bai, Z. and Silverstein, J. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd edition. Springer, New York.

Baik, J., Ben Arous, G. and Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, **33**, 1643–1697.

Baik, J. and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.*, **97**, 1382–1408.

Bakhshizadeh, M., Maleki, A. and de la Pena, V. H. (2020). Sharp concentration results for heavy-tailed distributions. Available at `https://arxiv.org/pdf/2003.13819.pdf`.

Barata, J. C. A. and Hussein, M. S. (2012). The Moore–Penrose pseudoinverse: A tutorial review of the theory. *Braz. J. Phys.*, **42**, 146–165.

Barbier, J., Dia, M., Macris, N., Krzakala, F., Lesieur, T. and Zdeborová, L. (2016). Mutual information for symmetric rank-one matrix estimation: a proof of the replica formula. In *Advances in Neural Information Processing Systems*, **29**, 424–432.

Barbier, J. and Krzakala, F. (2017). Approximate message-passing decoder and capacity achieving sparse superposition codes. *IEEE Trans. Inf. Theory*, **63**, 4894–4927.

Barbier, J., Krzakala, F., Macris, N., Miolane, L., and Zdeborová, L. (2019). Optimal errors and phase transitions in high-dimensional generalized linear models. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 5451–5460.

Barbier, J., Macris, N. and Rush, C. (2020). All-or-nothing statistical and computational phase transitions in sparse spiked matrix estimation. Available at https://arxiv.org/pdf/2006.07971.pdf.

Bartlett, P. L., Long, P. M., Lugosi, G. and Tsigler, A. (2020). Benign overfitting in linear regression. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 30063–30070.

Bayati, M., Lelarge, M., Montanari, A. (2015). Universality in polytope phase transitions and message passing algorithms. *Ann. Appl. Probab.*, **25**, 753–822.

Bayati, M. and Montanari, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inf. Theory*, **57**, 764–785.

Bayati, M. and Montanari, A. (2012). The LASSO risk for Gaussian matrices. *IEEE Trans. Inf. Theory*, **58**, 1997–2017.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci*, **2**, 183–202.

Belkin, M., Hsu, D., Ma, S. and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 15849–15854.

Belkin, M., Hsu, D. and Xu, J. (2020). Two models of double descent for weak features. *SIAM J. Math. Data Sci.*, **2**, 1167–1180.

Bellec, P. C., Lecué, G. and Tsybakov, A. B. (2018). SLOPE meets LASSO: improved oracle bounds and optimality. *Ann. Statist.*, **46**, 3603–3642.

Benaych-Georges, F. and Nadakuditi, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.*, **227**, 494–521.

Berthier, R., Montanari, A. and Nguyen, P.-M. (2020). State evolution for approximate message passing with non-separable functions. *Inf. Inference*, **9**, 33–79.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.*, **3**, 993–1022.

Bogdan, M., van den Berg, E., Sabatti, C., Su, W. and Candès, E. (2015). SLOPE—Adaptive variable selection via convex optimization. *Ann. Appl. Stat.*, **9**, 1103–1140.

Bolthausen, E. (2014). An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model. *Comm. Math. Phys.*, **325**, 333–366.

Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press, Oxford.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, **3**, 1–122.

Brown, L. D. and Purves, R. (1973). Measurable selections of extrema. *Ann. Statist.*, **1**, 902–912.

Bu, Z., Klusowski, J., Rush, C. and Su, W. (2021). Algorithmic analysis and statistical estimation of SLOPE via approximate message passing. *IEEE Trans. Inf. Theory*, **67**, 506–537.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer, Berlin.

Çakmak, B. and Opper, M. (2019). Memory-free dynamics for the Thouless–Anderson–Palmer equations of Ising models with arbitrary rotation-invariant ensembles of random coupling matrices. *Phys. Rev. E*, **99**, 062140.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.*, **9**, 717–772.

Candès, E. J. and Sur, P. (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *Ann. Statist.*, **48**, 27–42.

Capitaine, M., Donati-Martin, C. and Féral, D. (2009). The largest eigenvalues of finite rank deformation of large Wigner matrices: convergence and nonuniversality of the fluctuations. *Ann. Probab.*, **37**, 1–47.

Celentano, M. and Montanari, A. (2022). Fundamental barriers to high-dimensional regression with convex penalties. *Ann. Statist.*, to appear.

Celentano, M., Montanari, A. and Wu, Y. (2020). The estimation error of general first order methods. *Proc. Mach. Learn. Res.*, **125**, 1–64.

Chen, W-K. and Lam, W-K. (2021). Universality of approximate message passing algorithms. *Electron. J. Probab.*, **26**, 1–44.

Dar, Y., Muthukumar, V. and Baraniuk, R. (2021). A farewell to the bias-variance tradeoff? An overview of the theory of overparameterized machine learning. Available at https://arxiv.org/pdf/2109.02355.pdf.

d'Ascoli, S., Refinetti, M., Biroli, G. and Krzakala, F. (2020). Double trouble in double descent: bias and variance(s) in the lazy regime. *Proc. Mach. Learn. Res.*, **119**, 2280–2290.

Deng, Z., Kammoun, A. and Thrampoulidis, C. (2019). A model of double descent for high-dimensional binary linear classification. Available at https://arxiv.org/pdf/1911.05822.pdf.

Deshpande, Y., Abbe, E. and Montanari, A. (2016). Asymptotic mutual information for the balanced binary stochastic block model. *Inf. Inference*, **6**, 125–170.

Deshpande, Y. and Montanari, A. (2014). Information-theoretically optimal sparse PCA. In *2014 IEEE International Symposium on Information Theory*, pp. 2197–2201.

Deshpande, Y. and Montanari, A. (2015). Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time. *Found. Comput. Math.*, **15**, 1069–1128.

Donoho, D. L., Javanmard, A. and Montanari, A. (2013). Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Trans. Inf. Theory*, **59**, 7434–7464.

Donoho, D. L. and Johnstone, I. M. (1994). Minimax risk over l_p balls for l_q error. *Prob. Theory Related Fields*, **99**, 277–303.

Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.*, **26**, 879–921.

Donoho, D. L., Maleki, A. and Montanari, A. (2009). Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 18914–18919.

Donoho, D. and Montanari, A. (2015). Variance breakdown of Huber (M)-estimators: $n/p \to m \in (1, \infty)$. Available at https://arxiv.org/pdf/1503.02106.pdf.

Donoho, D. and Montanari, A. (2016). High dimensional robust M-estimation: asymptotic variance via approximate message passing. *Probab. Theory Related Fields*, **166**, 935–969.

Dudley, R. M. (2002). *Real Analysis and Probability*, 2nd edition. Cambridge University Press, Cambridge.

Dümbgen, L., Samworth, R. and Schuhmacher, D. (2011). Approximation by log-concave distributions, with applications to regression. *Ann. Statist.*, **39**, 702–730.

Dümbgen, L., Samworth, R. J. and Wellner, J. A. (2021). Bounding distributional errors via density ratios. *Bernoulli*, **27**, 818–852.

Efron, B. (2011). Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.*, **106**, 1602–1614.

El Alaoui, A, Ramdas, A., Krzakala, F., Zdeborová and Jordan, M. I. (2018). Decoding from pooled data: phase transitions of message passing. *IEEE Trans. Inf. Theory*, **65**, 572–585.

Emami, M., Sahraee-Ardakan, M., Pandit, P., Rangan, S. and Fletcher, A. K. (2020). Generalization error of generalized linear models in high dimensions. *Proc. Mach. Learn. Res.*, **119**, 2892–2901.

Fan, Z. (2022). Approximate message passing algorithms for rotationally invariant matrices. *Ann. Statist.*, to appear.

Federer, H. (1996). *Geometric Measure Theory.* Springer–Verlag, New York.

Féral, D. and Péché, S. (2007). The largest eigenvalue of rank one deformation of large Wigner matrices. *Comm. Math. Phys.*, **272**, 185–228.

Fletcher, A. K. and Rangan, S. (2014). Scalable inference for neuronal connectivity from calcium imaging. In *Advances in Neural Information Processing Systems*, **27**, 2843–2851.

Fletcher, A. K., Rangan, S. and Schniter, P. (2018). Inference in deep networks in high dimensions. In *2018 IEEE International Symposium on Information Theory*, pp. 1884–1888.

Fourdrinier, D., Strawderman, W. E. and Wells, M. T. (2018). *Shrinkage Estimation.* Springer, New York.

Gataric, M., Wang, T. and Samworth, R. J. (2020). Sparse principal component analysis via axis-aligned random projections. *J. Roy. Statist. Soc., Ser B*, **82**, 329–359.

Geiger, M., Jacot, A., Spigler, S., Gabriel, F., Sagun, L., d'Ascoli, S., Biroli, G., Hongler, C. and Wyart, M. (2019) Scaling description of generalization with number of parameters in deep learning. Available at https://arxiv.org/pdf/1901.01608.pdf.

Gerbelot, C., Abbara, A. and Krzakala, F. (2020a). Asymptotic errors for convex penalized linear regression beyond Gaussian matrices. *Proc. Mach. Learn. Res.*, **125**, 1682–1713.

Gerbelot, C., Abbara, A. and Krzakala, F. (2020b). Asymptotic errors for teacher-student convex generalized linear models (or: how to prove Kabashima's replica formula). Available at https://arxiv.org/pdf/2006.06581.pdf.

Gordon, L. (1994). A stochastic approach to the gamma function. *Am. Math. Mon.*, **101**, 858–865.

Guo, D. and Verdú, S. (2005). Randomly spread CDMA: Asymptotics via statistical physics. *IEEE Trans. Inf. Theory*, **51**, 1983–2010.

Han, Q. (2022). Noisy linear inverse problems under convex constraints: exact risk asymptotics in high dimensions. Available at https://arxiv.org/pdf/2201.08435.pdf.

Hastie, T., Montanari, A., Rosset, S. and Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Ann. Statist.*, to appear.

Hopkins, S. B., Shi, J. and Steurer, D. (2015). Tensor principal component analysis via sum-of-square proofs. *Proc. Mach. Learn. Res.*, **40**, 956–1006.

Hsu, P. L. and Robbins, H. (1947). Complete convergence and the law of large numbers. *Proc. Natl. Acad. Sci. U.S.A.*, **33**, 25–31.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, **35**, 73–101.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.*, **1**, 799–821.

Huber, P. J. and Ronchetti, E. (2009). *Robust Statistics*, 2nd edition. Wiley, New York.

Javanmard, A. and Montanari, A. (2013). State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Inf. Inference*, **2**, 115–144.

Jeon, C., Ghods, R., Maleki, A. and Studer, C. (2015). Optimality of large MIMO detection via approximate message passing. In *2015 IEEE International Symposium on Information Theory*, pp. 1227–1231.

Johnstone, I. M. (2006). High Dimensional Statistical Inference and Random Matrices. In *Proceedings of the International Congress of Mathematicians, Madrid 2006*, pp. 307–333.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.*, **104**, 682–693.

Johnstone, I. M. and Paul, D. (2018). PCA in high dimensions: an orientation. *Proc. IEEE*, **106**, 1277–1292.

Jolliffe, I. T., Trendafilov, N. T. and Uddin, M. (2003). A modified principal component technique based on the LASSO. *J. Comput. Graph. Statist.*, **12**, 531–547.

Kabashima, Y., Krzakala, F., Mézard, M., Sakata, A. and Zdeborová, L. (2016). Phase transitions and sample complexity in Bayes optimal matrix factorization. *IEEE Trans. Inf. Theory*, **62**, 4228–4265.

Kabashima, Y. and Vehkaperä, M. (2014). Signal recovery using expectation consistent approximation for linear observations. In *2014 IEEE International Symposium on Information Theory*, pp. 226–230.

Kallenberg, O. (1997). *Foundations of Modern Probability.* Springer–Verlag, New York.

Kini, G. R. and Thrampoulidis, C. (2020). Analytic study of double descent in binary classification: the impact of loss. In *2020 IEEE International Symposium on Information Theory*, pp. 2527–2532.

Knowles, A. and Yin, J. (2013). The isotropic semicircle law and deformation of Wigner matrices. *Comm. Pure Appl. Math.*, **66**, 1663–1749.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques.* MIT Press, Cambridge, Massachusetts.

Krzakala, F., Mézard, M., Sausset, F., Sun, Y. and Zdeborová, L. (2012). Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *J. Stat. Mech. Theory Exp.*, P08009.

Kuchibhotla, A. and Chakrabortty A. (2018). Moving beyond sub-Gaussianity in high-dimensional statistics: applications in covariance estimation and linear regression. Available at https://arxiv.org/pdf/1804.02605.pdf.

Lelarge, M. and Miolane, L. (2019). Fundamental limits of symmetric low-rank matrix estimation. *Probab. Theory Related Fields*, **173**, 859–929.

Lesieur, T., Krzakala, F. and Zdeborová, L. (2017). Constrained low-rank matrix estimation: phase transitions, approximate message passing and applications. *J. Stat. Mech. Theory Exp.*, 073403.

Li, Y. and Wei, Y. (2021). Minimum ℓ_1-norm interpolators: precise asymptotics and multiple descent. Available at https://arxiv.org/pdf/2110.09502.pdf.

Liang, T. and Rakhlin, A. (2020). Just interpolate: kernel 'ridgeless' regression can generalize. *Ann. Statist.*, **48**, 1329–1347.

Liang, T. and Sur, P. (2022). A precise high-dimensional asymptotic theory for boosting and minimum-ℓ_1-norm interpolated classifiers. *Ann. Statist.*, to appear.

Liu, L., Huang, S. and Kurkoski, B. M. (2021). Memory approximate message passing. In *2021 IEEE International Symposium on Information Theory*, pp. 1379–1384.

Ma, J. and Ping, L. (2017). Orthogonal AMP. *IEEE Access*, **5**, 2020–2033.

Ma, J., Xu, J. and Maleki, A. (2019). Optimization-based AMP for phase retrieval: the impact of initialization and ℓ_2 regularization. *IEEE Trans. Inf. Theory*, **65**, 3600–3629.

Ma, J., Xu, J. and Maleki, A. (2021). Impact of the sensing spectrum on signal recovery in generalized linear models. Available at https://arxiv.org/pdf/2111.03237.pdf.

Ma, Y., Rush, C. and Baron, D. (2019). Analysis of approximate message passing with non-separable denoisers and Markov random field priors. *IEEE Trans. Inf. Theory*, **65**, 7367–7389.

Ma, Z. and Wu, Y. (2015). Computational barriers in minimax submatrix detection. *Ann. Statist.*, **43**, 1089–1116.

Manoel, A., Krzakala, F., Mézard, M. and Zdeborová, L. (2017). Multi-layer generalized linear estimation. In *2017 IEEE International Symposium on Information Theory*, pp. 2098–2102.

Matsushita, R. and Tanaka, T. (2013). Low-rank matrix reconstruction and clustering via approximate message passing. In *Advances in Neural Information Processing Systems*, **26**, 917–925.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition. Chapman & Hall/CRC, Boca Raton.

Mehta, M. L. (2004). *Random Matrices*, 3rd edition. Elsevier, San Diego.

Mei, S. and Montanari, A. (2020). The generalization error of random features regression: precise asymptotics and double descent curve. *Comm. Pure Appl. Math.*, **75**, 667–766.

Metzler, C., Mousavi, A. and Baraniuk, R. (2017). Learned D-AMP: Principled neural network based compressive image recovery. In *Advances in Neural Information Processing Systems*, **30**, 1772–1783.

Mézard, M. and Montanari, M. (2009). *Information, Physics, and Computation.* Oxford University Press, Oxford.

Mézard, M., Parisi, G., Virasoro, M. A. (1987). *Spin Glass Theory and Beyond.* World Scientific Lecture Notes in Physics, **9**.

Miolane, L. and Montanari, A. (2021). The distribution of the Lasso: uniform control over sparse balls and adaptive parameter tuning. *Ann. Statist.*, **49**, 2313–2335.

Mondelli, M., Thrampoulidis, C. and Venkataramanan, R. (2021). Optimal combination of linear and spectral estimators for generalized linear models. *Found. Comput. Math.*, 2021.

Mondelli, M. and Venkataramanan, R. (2020). Approximate message passing with spectral initialization for generalized linear models. *Proc. Mach. Learn. Res.*, **130**, 397–405.

Mondelli, M. and Venkataramanan, R. (2021). PCA initialization for approximate message passing in rotationally invariant models. Available at https://arxiv.org/pdf/2106.02356.pdf.

Montanari, A. (2012). Graphical Models Concepts in Compressed Sensing. In *Compressed Sensing: Theory and Applications* (Y. Eldar and G. Kutyniok, eds.). Cambridge University Press, Cambridge.

Montanari, A. and Richard, E. (2014). A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems*, **27**, 2897–2905.

Montanari, A. and Richard, E. (2016). Non-negative principal component analysis: Message passing algorithms and sharp asymptotics. *IEEE Trans. Inf. Theory*, **62**, 1458–1484.

Montanari, A. and Venkataramanan, R. (2021). Estimation of low-rank matrices via approximate message passing. *Ann. Statist.*, **49**, 321–345.

Mousavi, A., Maleki, A., Baraniuk, R. G. (2018). Consistent parameter estimation for LASSO and approximate message passing. *Ann. Statist.*, **46**, 119–148.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B. and Sutskever, I. (2021) Deep double descent: where bigger models and more data hurt. *J. Stat. Mech. Theory Exp.*, 124003.

Opper, M., Çakmak, B. and Winther, O. (2016). A theory of solving TAP equations for Ising models with general invariant random matrices. *J. Phys. A.*, **49**, 114002.

Opper, M. and Winther, O. (2005). Expectation consistent approximate inference. *J. Mach. Learn. Res.*, **6**, 2177–2204.

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference: From a Neo-Fisherian Perspective.* World Scientific, Singapore.

Panaretos, V. M. and Zemel, Y. (2020). *An Invitation to Statistics in Wasserstein Space.* Springer–Verlag, New York.

Pandit, P., Sahraee-Ardakan, M., Rangan, S., Schniter, P. and Fletcher, A. K. (2020). Inference with deep generative priors in high dimensions. *IEEE J. Sel. Areas Inf. Theory*, **1**, 336–347.

Parikh, N. and Boyd, S. (2013). Proximal algorithms. *Found. Trends Optim.*, **1**, 123–231.

Parker, J. T., Schniter, P. and Cevher, V. (2014a). Bilinear generalized approximate message passing—Part I: Derivation. *IEEE Trans. Signal Process.*, **62**, 5839–5853.

Parker, J. T., Schniter, P. and Cevher, V. (2014b). Bilinear generalized approximate message passing—Part II: Applications. *IEEE Trans. Signal Process.*, **62**, 5854–5867.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica*, **17**, 1617–1642.

Peng, M. (2012). Eigenvalues of deformed random matrices. Available at https://arxiv.org/pdf/1205.0572.pdf.

Perry, A., Wein, A. S., Bandeira, A. S. and Moitra, A. (2018). Optimality and sub-optimality of PCA I: Spiked random matrix models. *Ann. Statist.*, **46**, 2416–2451.

Portnoy, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Statist.*, **12**, 1298–1309.

Portnoy, S. (1985). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large; II. Normal approximation. *Ann. Statist.*, **13**, 1403–1417.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.*, **16**, 356–366.

Prékopa, A. (1980). Logarithmic concave measures and related topics. In *Stochastic Programming (Proc. Internat. Conf., Univ. Oxford, Oxford, 1974, M. A. H. Dempster ed.)*, pp. 63–82. Academic Press, London.

Rangan, S. (2011). Generalized approximate message passing for estimation with random linear mixing. In *2011 IEEE International Symposium on Information Theory*, pp. 2168–2172.

Rangan, S. and Fletcher, A. K. (2012). Iterative estimation of constrained rank-one matrices in noise. In *2012 IEEE International Symposium on Information Theory*, pp. 1246–1250.

Rangan, S. and Fletcher, A. K. (2018). Iterative reconstruction of rank-one matrices in noise. *Inf. Inference*, **7**, 1246–1250.

Rangan, S., Fletcher, A. K. and Goyal, V. K. (2009). Asymptotic analysis of MAP estimation via the replica method and applications to compressed sensing. In *Advances in Neural Information Processing Systems*, **22**, 1545–1553.

Rangan, S., Schniter, P., Fletcher, A. K. and Sarkar, S. (2019a). On the convergence of approximate message passing with arbitrary matrices. *IEEE Trans. Inf. Theory*, **65**, 5339–5351.

Rangan, S., Schniter, P. and Fletcher, A. K. (2019b). Vector approximate message passing. *IEEE Trans. Inf. Theory*, **65**, 6664–6684.

Rangan, S., Schniter, P., Riegler, E., Fletcher, A. K. and Cevher, V. (2016). Fixed points of generalized approximate message passing with arbitrary matrices. *IEEE Trans. Inf. Theory*, **62**, 7464–7474.

Reeves, G. and Pfister, H. D. (2019). The replica-symmetric prediction for random linear estimation with Gaussian matrices is exact. *IEEE Trans. Inf. Theory*, **65**, 2252–2283.

Robbins, H. (1956). An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. Math. Statist. Prob.*, **1**, 157–163.

Rockafellar, R. T. (1997). *Convex Analysis.* Princeton University Press, Princeton.

Rush, C., Greig, A. and Venkataramanan, R. (2017). Capacity-achieving sparse superposition codes via approximate message passing decoding. *IEEE Trans. Inf. Theory*, **63**, 1476–1500.

Rush, C. and Venkataramanan, R. (2018). Finite sample analysis of approximate message passing algorithms. *IEEE Trans. Inf. Theory*, **64**, 7264–7286.

Schniter, P. (2011). A message-passing receiver for BICM-OFDM over unknown clustered-sparse channels. *IEEE J. Sel. Top. Signal Process.*, **5**, 1462–1474.

Schniter, P. (2020). A simple derivation of AMP and its state evolution via first-order cancellation. *IEEE Trans. Signal Process.*, **68**, 4283–4292.

Schniter, P. and Rangan, S. (2014). Compressive phase retrieval via generalized approximate message passing. *IEEE Trans. Signal Process.*, **63**, 1043–1055.

Schniter, P., Rangan, S. and Fletcher, A. K. (2016). Vector approximate message passing for the generalized linear model. In *50th Asilomar Conference on Signals, Systems and Computers*, pp. 1525–1529.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley, New York.

Su, W., Bogdan, M. and Candès, E. (2017). False discoveries occur early on the LASSO path. *Ann. Statist.*, **45**, 2133–2150.

Su, W. and Candès, E. (2016). SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.*, **44**, 1038–1068.

Su, X. and Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Adv. Artif. Intelligence*, Volume 2009, 1–19.

Sur, P. and Candès, E. J. (2019a). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 14516–14525.

Sur, P. and Candès, E. J. (2019b). Additional supplementary materials for 'A modern maximum-likelihood theory for high-dimensional logistic regression'. Available at https://sites.fas.harvard.edu/~prs499/papers/proofs_LogisticAMP.pdf.

Sur, P., Chen, Y. and Candès, E. J. (2017). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probab. Theory Related Fields*, **175**, 487–558.

Takeuchi, K. (2020). Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements. *IEEE Trans. Inf. Theory*, **66**, 368–386.

Takeuchi, K. (2021a). Bayes-optimal convolutional AMP. *IEEE Trans. Inf. Theory*, **67**, 4405–4428.

Takeuchi, K. (2021b). On the convergence of Orthogonal/Vector AMP: long-memory message-passing strategy. Available at https://arxiv.org/pdf/2111.05522.pdf.

Talagrand, M. (2011). *Mean Field Models for Spin Glasses, Vol I: Basic Examples.* Springer, New York.

Tanaka, T. (2002). A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors. *IEEE Trans. Inf. Theory*, **48**, 2888–2910.

Thrampoulidis, C., Abbasi, E. and Hassibi, B. (2018). Precise error analysis of regularized M-estimators in high dimensions. *IEEE Trans. Inf. Theory*, **64**, 5592–5628.

Thrampoulidis, C., Oymak, S. and Hassibi, B. (2015). Regularized linear regression: a precise analysis of the estimation error. *Proc. Mach. Learn. Res.*, **40**, 1683–1709.

Tian, F., Liu, L. and Chen, X. (2021). Generalized memory approximate message passing. Available at https://arxiv.org/pdf/2110.06069.pdf.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc., Ser. B*, **58**, 267–288.

Tramel, E. W., Kumar, S., Giurgiu, A. and Montanari, A. (2014). Statistical estimation: from denoising to sparse regression and hidden cliques. Available at https://arxiv.org/pdf/1409.5557.pdf.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation.* Springer–Verlag, New York.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press, Cambridge.

Venkataramanan, R., Kögler, K. and Mondelli, M. (2021). Estimation in rotationally invariant generalized linear models via approximate message passing. Available at https://arxiv.org/pdf/2112.04330.pdf.

Vila, J., Schniter, P. and Meola, J. (2015). Hyperspectral unmixing via turbo bilinear approximate message passing. *IEEE Trans. Comput. Imaging*, **1**, 143–158.

Villani, C. (2003). *Topics in Optimal Transportation. Graduate Studies in Mathematics.* American Mathematical Society, Providence, RI.

Villani, C. (2009). *Optimal Transport, Old and New.* Springer–Verlag, New York.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statist. Comput.*, **17**, 395–416.

Vu, V. Q. and Lei, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.*, **41**, 2905–2947.

Wang, T., Berthet, Q. and Samworth, R. J. (2016). Statistical and computational trade-offs in estimation of sparse principal components. *Ann. Statist.*, **44**, 1896–1930.

Wein, A. S., El Alaoui, A. and Moore, C. (2019). The Kikuchi hierarchy and tensor PCA. In *2019 IEEE Annual Symposium on Foundations of Computer Science*, pp. 1446–1468.

Yang, G. (2019). Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. Available at https://arxiv.org/pdf/1902.04760.pdf.

Zdeborová, L. and Krzakala, F. (2016). Statistical physics of inference: thresholds and algorithms. *Adv. Phys.*, **65**, 453–552.

Zhong, X., Wang, T. and Fan, Z. (2021). Approximate message passing for orthogonally invariant ensembles: Multivariate non-linearities and spectral initialization. Available at https://arxiv.org/pdf/2110.02318.pdf.

Zhu, Z., Wang, T. and Samworth, R. J. (2019). High-dimensional principal component analysis with heterogeneous missingness. Available at https://arxiv.org/pdf/1906.12125.pdf.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.*, **15**, 265–286.